



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

**EVALUATING AND IMPROVING THE SAMA
(SEGMENTATION ANALYSIS AND MARKET
ASSESSMENT) RECRUITING MODEL**

by

William N. Marmion

June 2015

Thesis Advisor:
Co-Advisor:

Lyn Whitaker
Jonathan K. Alt

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE June 2015	3. REPORT TYPE AND DATES COVERED Master's Thesis	
4. TITLE AND SUBTITLE EVALUATING AND IMPROVING THE SAMA (SEGMENTATION ANALYSIS AND MARKET ASSESSMENT) RECRUITING MODEL			5. FUNDING NUMBERS	
6. AUTHOR(S) William N. Marmion			8. PERFORMING ORGANIZATION REPORT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) United States Army Recruiting Command Fort Knox, KY				
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ____N/A____.				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited			12b. DISTRIBUTION CODE	
13. ABSTRACT (maximum 200 words) Military recruiting for an all-volunteer force requires deliberate planning and market analysis in order to achieve prescribed recruiting goals. United States Army Recruiting Command (USAREC) leaders and planners leverage existing tools and technology to measure performance and potential within their areas of responsibility. One of the tools used by USAREC is the Segmentation Analysis and Market Assessment (SAMA) tool. This tool calculates recruitment potential of recruiting centers using a four-year weighted performance average within customized Army market segments. An analysis of the current SAMA model shows it overestimates production potential for 96% of centers, leading decision makers to set unrealistic goals for their organizations. The use of additional factors and alternative modeling approaches, Least Squared Regression, and Neural Networks, results in models with greater predictive power, while avoiding overestimation. The statistical models developed in this thesis match the predictive power of the current SAMA methodology while overestimating average potential by only 3.8%. More precise modeling tools will improve USAREC's ability to more effectively plan recruiting operations and allocate resources.				
14. SUBJECT TERMS Recruiting			15. NUMBER OF PAGES 87	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

**EVALUATING AND IMPROVING THE SAMA (SEGMENTATION ANALYSIS
AND MARKET ASSESSMENT) RECRUITING MODEL**

William N. Marmion
Captain, United States Army
B.S., United States Military Academy, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
June 2015**

Author: William N. Marmion

Approved by: Lyn Whitaker
Thesis Advisor

Jonathan K. Alt
Co-Advisor

Robert F. Dell
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Military recruiting for an all-volunteer force requires deliberate planning and market analysis in order to achieve prescribed recruiting goals. United States Army Recruiting Command (USAREC) leaders and planners leverage existing tools and technology to measure performance and potential within their areas of responsibility. One of the tools used by USAREC is the Segmentation Analysis and Market Assessment (SAMA) tool. This tool calculates recruitment potential of recruiting centers using a four-year weighted performance average within customized Army market segments. An analysis of the current SAMA model shows it overestimates production potential for 96% of centers, leading decision makers to set unrealistic goals for their organizations. The use of additional factors and alternative modeling approaches, Least Squared Regression, and Neural Networks, results in models with greater predictive power, while avoiding overestimation. The statistical models developed in this thesis match the predictive power of the current SAMA methodology while overestimating average potential by only 3.8%. More precise modeling tools will improve USAREC's ability to more effectively plan recruiting operations and allocate resources.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I.	INTRODUCTION.....	1
A.	THE UNITED STATES ARMY RECRUITING COMMAND.....	1
B.	RECRUITING OPERATIONS.....	2
1.	Recruiting Center Organization.....	3
2.	Market Information and Processing.....	4
3.	Segmentation Analysis and Market Assessment tool.....	4
C.	OVERVIEW.....	5
II.	BACKGROUND.....	7
A.	USAREC G2 AND THE MISSION PROCESS.....	7
B.	CURRENT SAMA METHODOLOGY AND SEGMENTATION.....	9
C.	PREVIOUS EXPLORATION.....	12
D.	OTHER RELATED WORKS.....	15
III.	DATA COLLECTION AND PREPARATION.....	17
A.	SAMA SPECIFIC ANALYSIS.....	17
B.	ADDITIONAL FACTORS.....	17
C.	SAMA CALCULATOR AND CONSOLIDATED DATASET.....	18
IV.	ANALYSIS OF CURRENT SAMA CALCULATIONS.....	21
A.	CURRENT SAMA CALCULATIONS VS. 2014 ACHIEVEMENT.....	22
B.	PRIZM NE SAMA CALCULATIONS VERSUS 2014 ACHIEVEMENT.....	23
V.	MODELING.....	27
A.	PRIZM NE CONSOLIDATION SCORES AS SIMPLIFIED FACTORS.....	27
1.	CORE Score.....	28
2.	SOCIAL Score and LIFE Score.....	30
B.	EXAMINATION OF DATA.....	32
C.	MULTIPLE LINEAR REGRESSION MODEL WITH PREVIOUS PERFORMANCE.....	33
1.	Formulating and Fitting the Model.....	34
2.	Checking Assumptions and Validation.....	36
D.	MULTIPLE LINEAR REGRESSION WITHOUT PREVIOUS PERFORMANCE.....	40
E.	NEURAL NETWORK MODEL.....	41
1.	Formulating and Fitting the Model.....	42
2.	Model Selection.....	45
VI.	MODEL COMPARISON AND CONCLUSION.....	49
A.	MODEL COMPARISON.....	49
B.	CONCLUSION.....	51
C.	FUTURE WORK.....	51
	APPENDIX A. SAMA POTENTIAL CALCULATIONS.....	53

A.	SEGMENTATION AND FOUR-YEAR WEIGHTED AVERAGE	53
B.	PENETRATION RATES.....	53
APPENDIX B.	SAMPLE VBA CONSOLIDATION CODE	55
APPENDIX C.	DATA EXAMINATION.....	57
APPENDIX D.	TRANSFORMED PREDICTOR VALUE PLOTS	61
LIST OF REFERENCES	63
INITIAL DISTRIBUTION LIST	65

LIST OF FIGURES

Figure 1.	The United States Army Recruiting Command Brigades and Battalions (from USAREC, 2013).	2
Figure 2.	Organization of a Recruiting Center (from USAREC, 2012).	3
Figure 3.	Missioning Model (from Devin, 2015)	8
Figure 4.	SAMA Reports Overview (from USAREC G2, 2012).	10
Figure 5.	Tactical Segmentation Market Report Example (from USAREC G2, 2012).	11
Figure 6.	SAMA Potential to Actual for 843 Centers (from Devin, SAMA Methodology Validation, 2014).	14
Figure 7.	SAMA Index Average by Brigade including individual centers	21
Figure 8.	Simple Linear Regression of 2014 Contract Achievement by SAMA Calculated Potential (from Army Custom Segments)	22
Figure 9.	SAMA PRIZM Index Average by Brigade including individual centers	24
Figure 10.	Simple Linear Regression of 2014 Contract Achievement by SAMA PRIZM NE Calculated Potential	25
Figure 11.	Four Year (2010–2014) Weighted Average of Total Contracts per PRIZM NE Segment	28
Figure 12.	PRIZM NE Social Groups (from Claritas, 2015)	30
Figure 13.	PRIZM NE Lifestage Groups (from Claritas, 2015)	31
Figure 14.	Initial Linear Regression Model Actual by Predicted Plot and Summary of Fit	35
Figure 15.	Box-Cox Transformation Plot for Linear Regression Model	36
Figure 16.	Linear Regression Model with Power Transformed Response	37
Figure 17.	Residual Values against the Predicted Values plot, Linear Regression Model	38
Figure 18.	Residuals plot (with accompanying Normal QQ line plot), Linear Model	39
Figure 19.	Diagram of a Neural Network Architecture (from Yu-Wei, 2015)	42
Figure 20.	JMP 11 PRO Neural Network Launch Window	44
Figure 21.	Plots of the response (2014 Number of Contracts) QMA, Recruiters, and Unemployment Rate	57
Figure 22.	Plots of the response (2014 Number of Contracts) against weighted four-year average enlistments, weighted four-year average enlistments for all services, and CORE Score	57
Figure 23.	Plots of the response (2014 Number of Contracts) against SOCIAL Score, LIFE Score, and square mileage of center area of responsibility	58
Figure 24.	Plots of the response (2014 Number of Contracts) against distance (miles) from center to nearest MEPS, and driving time from center to nearest MEPS	58
Figure 25.	Scatterplot Matrix	59
Figure 26.	Transformed Predictor Regression Plots–Previous Performance	61
Figure 27.	Transformed Predictor Regression Plots–No Previous Performance	62

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1.	Clarksville Company Data for 2013, Company and Centers (from Devin, 2014)	12
Table 2.	SAMA Potential Calculations for Clarksville Company as of September 2014 (from Devin, 2014)	13
Table 3.	Sample of Consolidated Data for Additional Factors	18
Table 4.	CORE PRIZM NE Segments.....	29
Table 5.	Neural Network Model Results	46
Table 6.	Neural Network Model Narrowed Results	47
Table 7.	Model Comparison.....	50
Table 8.	Correlation Matrix	60

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

4yr_Wtd_Avg	Four Year Weighted Average
ACS	Army Custom Segments
AIC	Akaike Information Criterion
AR	Army Reserve
ASVAB	Armed Services Vocational Aptitude Battery
BDE	Brigade
BIC	Bayesian Information Criterion
CG	Commanding General
CO	Company
COA	Course of Action
CTR	Center
DOD	Department of Defense
GA	Graduate Alpha
GED	General Educational Development
IET	Initial Entry Training
MEPS	Military Entrance Processing Station
MRB	Medical Recruiting Brigade
OTH	Other
Pen Rate	Penetration Rate
PenRate _{RS}	Center Penetration Rate
PenRate _{RTC}	Company Penetration Rate
PotentialPenRate _{TS}	Potential Penetration Rate by Tactical Segment
PRIZM NE	Potential Rating Index for Zip Markets, New Evolution
QMA	Qualified Military Available
QQ	Quantile-Quantile
QTB	Quarterly Training Brief
RA	Regular Army
RBN	Recruiting Battalion
RMSE	Root-Mean-Squared Error
RS	Recruiting Station

RSID	Recruiting Station Identification
RTC	Recruiting Company
SA	Senior Alpha
SAMA	Segmentation Analysis and Market Assessment
SSE	Sum of Squared Errors
TS	Tactical Segment
USAREC	United States Army Recruiting Command
VBA	Visual Basic for Applications
VIF	Variance Inflation Factor
YTD	Year-To-Date

EXECUTIVE SUMMARY

The United States Army Recruiting Command (USAREC) is the responsible U.S. Army organization for enlisting an all-volunteer Army for the service. The command is organized into five enlisted recruiting brigades and a medical recruiting brigade, and within each of these brigades are seven to eight battalions consisting of the recruiting companies and subordinate recruiting centers (CTR). The recently implemented concept of “recruiting centers” is a result of the first major organizational change within USAREC since its activation in 1964. These centers now operate using more efficient and team-oriented recruiting processes. These changes are critical due to the ongoing reduction in recruiting personnel and budgets. The ability to estimate an area’s recruitment potential is a necessity in planning for deliberate recruiting operations. One of the tools used by USAREC to perform these calculations is the Segmentation Analysis and Market Assessment (SAMA) tool (USAREC G2, 2012). This tool calculates the recruitment potential of recruiting centers using customized Army market segments. The purpose of this research is to validate the underlying SAMA model and propose an alternative methodology. This is important because more precise modeling tools will provide USAREC with the ability to more effectively plan recruiting operations and allocate resources.

The SAMA tool calculates the potential for a unit through comparison of its weighted four-year performance (number of enlistments) within each Army Tactical Segment with that of its next highest headquarter’s average for the segment. This provides good predictive power but leads to overestimations of potential in most cases. A limitation of this research is the lack of access to SAMA calculations since it is a real-time system available exclusively to USAREC personnel. However, we construct a SAMA calculator using performance data by segment over the past five years. A statistical analysis of the potential measurements shows that SAMA overestimates potential by at least 25% for 96% of USAREC centers. Fitting a simple linear regression using least squares for SAMA calculated potential against actual performance suggests close predictive power for SAMA despite the clear trend of over projecting potential.

As part of the data collection and cleaning for the analysis, we collect data sets identified by previous researchers as being relevant for measuring potential (Dereu, 1983). Additionally, we construct several summary scores based on the Claritas Potential Rating Index for Zip Markets, New Evolution (PRIZM NE) segmentation data to explore as additional factors. We construct three scores that implement a combination of performance, social, and lifestyle scores within the 66 PRIZM NE segments. Individual analysis of these factors and other factors results in several factor eliminations due to instances of correlation, with the following factors remaining for implementation in further modeling:

- Weighted four-year average enlistments
- Weighted four-year average enlistment for all services
- QMA (Qualified Military Available, aged 17–24 years)
- Recruiters (number of recruiters assigned to recruiting center)
- Unemployment Rate
- Driving time from center to nearest military processing station
- Score based on representation of high performing segments within a center
- Score based on representation of high performing social groups within a center
- Score based on representation of high performing lifestyle groups within a center

Regression is a very appropriate modeling technique, particularly since we are interested in predictions and forecasting a performance. We use multiple linear regression for our first two models with the number of 2014 enlistments for a center representing the dependent variable. We formulate models using the two direct previous performance factors and ones that omit them. Though previous performance is typically the most dominant predictive factor in determining future performance, it does not allow for changes in market trends, command influence, or policies within the organization. The resulting two models provide good potential predictive power and meet all of the required

modeling assumptions necessary for a valid model. Finally, we explore the use of artificial neural networks in building a predictive model. This model also does not use previous performance factors. Through validation using separate training, validation, and test data sets, we are able to generate a model with the right amount of complexity that predicts well without over-fitting the data.

Comparing the models, we recommend the first linear regression model, which uses a power transformation of the number of 2014 enlistments for a center and uses previous performance factors. This model should replace the current model because it only overestimates average potential by only 3.8%. We select this model because of its high R-squared score, its low root-mean-square error (RMSE), and its validity as an adequate model. USAREC personnel can easily update the model annually using the same methodology as outlined in the research using JMP software. Implementation of this model as part of the SAMA reports can lead to USAREC's ability to improve recruiting operations and allocation of resources.

References

- Dereu, J., & Robbin, J. (1983). *Application of geodemographics to the Army recruiting problem*. Fort Sheridan: Department of the Army.
- USAREC. (2012). *Recruiting center operations*. Fort Knox: United States Army Recruiting Command.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I'd like to thank my advisers, Dr. Lyn Whitaker and LTC Jon Alt, for their helpful expertise, guidance, and ability to assist me in understanding our research and leading me in developing a genuine interest in data analysis. Thanks to the USAREC G2 team (Mr. Michael Nelson, Mr. Mitch Stokan, and Major David Devin) for their great facilitation and continuous support. Thanks to my father, for setting a standard for hard work and dedication of a scholarly nature that I will never reach, but can thrive for during my own endeavors. Most importantly, to my wife Heather and my daughter Anya, thank you for providing me the strength and will to do my best, and rewarding me with your love every day.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. THE UNITED STATES ARMY RECRUITING COMMAND

The United States Army Recruiting Command (USAREC) was activated on October 1, 1964, with Fort Monroe serving as the headquarters (G-5 Public Affairs, USAREC, 2004). The U.S. Army has practiced forms of recruiting activities since 1822, but the reorganization and modernization of recruiting operations that we see today began in 1962, leading to USAREC's activation. The current USAREC mission is that from "1 October 2013 through September 2019, the Army (USAREC) will recruit professional, volunteer Soldiers; Soldiers capable of effectively executing operations in the Army's complex operating environment" (USAREC, 2013). USAREC is commanded by a Major General (2-star flag officer) and the current headquarters is located in Fort Knox, Kentucky. The command is organized into five brigades (BDE) and a medical recruiting brigade (MRB), each commanded by a Colonel. Figure 1 illustrates the regions of responsibility for each brigade. This illustration lists each brigade and the medical recruiting brigade along with their associated battalions. The number and the letter before each battalion are part of the designation code that USAREC uses to identify units. Within each of these battalions are recruiting companies and subordinate recruiting centers (CTR). The concept of "recruiting centers" is a result of the first major organizational change within USAREC since its activation in 1964. The approach "transforms and modernizes the legacy processes that have remained virtually unchanged since 1973" (USAREC, 2012). These recruiting centers now operate using more efficient and team oriented recruiting processes. These implementations are necessary due to the recent and ongoing reductions in recruiting personnel and budgets. The purpose of this research is to evaluate and improve the Segmentation Analysis and Market Assessment (SAMA) tool's performance. USAREC leaders and planners use SAMA to provide a standard methodology for battalions, companies, and centers to identify, prioritize, and target the various markets within their areas. This is important because more precise modeling tools will provide leaders and planners in USAREC the ability to more effectively plan recruiting operations and allocate resources.

Brigades and Battalions



Figure 1. The United States Army Recruiting Command Brigades and Battalions (from USAREC, 2013).

B. RECRUITING OPERATIONS

The Army assigns selected Soldiers in the conventional Army of the rank of Sergeant or Staff Sergeant to recruiting duty. Recruiting duty typically lasts for three years. These Soldiers do not possess previous military recruiting training or experience, but they attend the United States Army Recruiting and Retention School (now located in Fort Knox, Kentucky) for six weeks. Here, they learn decisive, supporting, and sustaining recruiting operations along with skills, techniques, and tools pivotal for the

successful recruiting of an all-volunteer force. Center Commanders, usually in the rank of Sergeant First Class, supervise recruiting centers. USAREC accepts these Center Commanders as full time recruiters following their regular recruiting duty (Military Occupation Skill 79-R).

1. Recruiting Center Organization

As of 2012, the recruiting center is led and supervised by a Center Commander, with an Assistant Center Commander serving as a second in command and an operations officer. The recruiting support team provides information on potential recruits to the engagement team as well as information on locations to conduct prospecting activities. The engagement team serves as the “face and voice of the Army in the community” (USAREC, 2012) and is the primary point of contact for potential recruits. Finally, the Future Soldier leader leads and manages those individuals who are awaiting their schedule date for Initial Entry Training. Figure 2 illustrates the organization of a recruiting center.

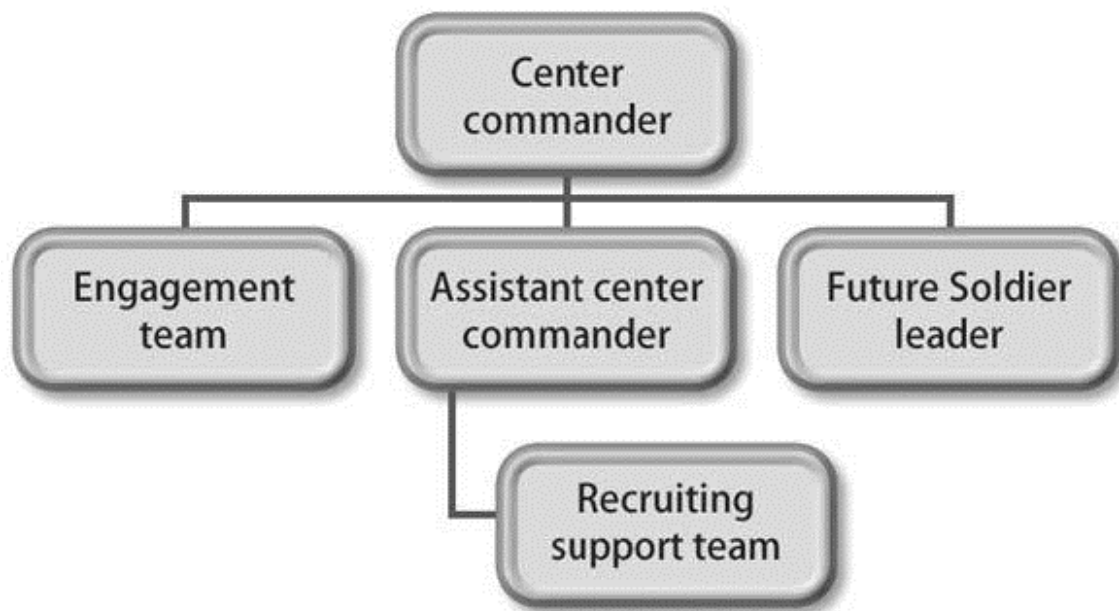


Figure 2. Organization of a Recruiting Center (from USAREC, 2012).

2. Market Information and Processing

In the past, individual recruiters received a monthly mission for which they were personally responsible. Though leaders still provided supervision and some direction, the individual recruiter was responsible for all of the steps of the recruiting process, from gathering market intelligence to shipping a future Soldier to training. Inefficiencies in this system led formal operational changes in 2009. Today, the Center Commanders receive the recruiting mission and are responsible for measuring performance and potential within their areas of responsibility. One of the tools used by leaders at all levels within USAREC is the SAMA tool.

3. Segmentation Analysis and Market Assessment tool

The SAMA tool is available through Report Management Zone, an interface native to all computer networks available to Army recruiters (USAREC G2, 2012). Along with a Market Assessment Report, the tool provides a Real Time Tactical Segmentation Report that calculates the “potential” enlistments (prediction) for each recruiting center. The model uses industry standard market analysis, applying aggregate regional statistics to set the performance objectives to a subset of the region (the individual zip codes) (M. Nelson, personal communication, December 16, 2014). SAMA currently uses the Army Custom Segments (ACS) to partition the population into 39 separate groups or tactical segments and then calculates penetration rates based upon weighted averages of recruiting rates over the previous four years. These penetration rates are used to calculate the potential of a recruiting center, company, or battalion through comparisons with adjacent and higher units. The model also gives the recruiting performance within each Tactical Segment. No other factors are accounted for in calculating these potentials other than previous performance. This sometimes leads to misleading results due to high recruiting rates in adjacent areas (M. Nelson, personal communication, December 16, 2014). The comparative method between center and company also leads to the overestimation of potential.

C. OVERVIEW

With a decrease in the number of recruiters and the transition to “Recruiting Operations,” accurate market analysis and predictions are critical for the leaderships’ planning process. Company level recruiting leadership will benefit from the evaluation of the current model used within SAMA and the identification of additional factors to include in the model. Alternative modeling approaches are also worth investigating since the current model regularly overestimates what is actually achieved by a center. Chapter II covers previous work explaining the current SAMA methodology and segmentation as well as provides an insight into previous related research. Chapter III describes the collection of data required to evaluate SAMA and to develop alternative models for predicting annual recruiting numbers for recruiting centers. Included are the details of the compilation, formatting, and coding as well as the data cleaning methodology. Chapter IV documents the analysis of the current SAMA methodology. Chapter V identifies other significant factors as well as the efforts in alternative modeling approaches using these new factors. Finally, Chapter VI summarizes the findings and results, presenting the findings of current methodologies as well as a comparison of the alternative models.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

A. USAREC G2 AND THE MISSION PROCESS

Before discussing SAMA, it is important to understand the role of the USAREC G2 and the missioning process. The USAREC G2 is the Commanding General's staff proponent for market intelligence and mission analysis and coordinates the positioning and missioning processes (USAREC, 2009):

- Alignment
- Positioning the Force
- Recruiter Requirement
- Recruiter Distribution
- Missioning the Force

The Commanding General tasks the USAREC G2 to analyze USAREC's accession mission in light of the existing Future Soldier posture so that the supporting contract mission can be developed. The contract mission is the number of individuals that must enlist and be placed in the Future Soldier Training Program each year. The time an individual spends in the Future Soldier Training Program varies but is typically between three months and a year. In 2014, USAREC began providing an annual assigned mission versus the previous assigned monthly missions (M. Stokan, personal communication, December 16, 2014). The Commanding General established this policy to give recruiting units understandable and predicable recruiting missions. The term "missioning" defines the process of assigning a distributed recruiting contract mission to the recruiting units that best enables the command to achieve the overall mission. The missioning process depends on the USAREC G2's market analysis, models, positioning of the force, and the information and analysis of subordinate recruiting leaders (USAREC, 2009). The USAREC headquarters allocates mission by four different factors:

- Brigade and battalion regions
- Component (Active Duty and Army Reserve)

- By required recruit quality standards (Armed Forces Qualification Test and education level):
 - i. Graduate Alpha (GA)–a high school graduate with good test scores (top 50%).
 - ii. Senior Alpha (SA)–a high school senior with good test scores.
 - iii. Other (OTH) category
 - High school graduates and seniors with lower test scores (31–49%)
 - General Educational Development (GED) applicant with good test scores
- By time to synchronize with Army systems such as Initial Entry Training (IET) seats and man-years for end-strength objectives

USAREC currently uses a weighted average of three elements to assign mission: a four-year weighted average of Army recruiting production, the Department of Defense (DOD) enlistment rate of individuals with good test scores (Grad-Senior Alphas), and a metric known as Qualified Military Available (QMA). QMA depicts the number of the population of an area aged 17–24 years. Figure 3 depicts USAREC’s courses of actions used in setting mission requirements.

Combined Enlistment Mission Considerations			
COAs	Description	Strengths	Weaknesses
COA 1 (60/0/40)	60% DoD GSA past production 40% QMA	Stable with reasonable accuracy for the past 4 years.	Focus on DoD past production is less sensitive to the Army potential and limits precise mission placement in an adverse recruiting environment
COA 2 (50/20/30)	50% DoD GSA past production 20% Army past production 30% QMA	Missions are robust as recruiting environment becomes more challenged by market forces.	Increases the total past production to 70% which is less sensitive to population shifts & emerging markets.

Figure 3. Missioning Model (from Devin, 2015)

Prior to July 2014, USAREC used Course of Action (COA) 1. This model weighted 60% DOD Grad-Senior Alpha production with 40% QMA. COA 2 weights

50% of DOD past production with 30% QMA and 20% of recruiting past production specific to the Army. Past production is weighted over the previous four years (40% 1st year, 30% 2nd year, 20% 3rd year, and 10% 4th year). Both methods have their strengths and weaknesses, but both use only QMA and previous production as factors when determining the mission distribution. Once this combined Active Duty and Army Reserve mission is determined, the missioning for the Army Reserve is calculated using a separate method. The Army Reserve mission is then subtracted from the combined mission to get the Active Duty mission. The SAMA tool only considers the Active Duty mission.

B. CURRENT SAMA METHODOLOGY AND SEGMENTATION

When USAREC personnel access the SAMA reports, real-time calculations of five sub-reports are displayed as outlined in Figure 4. The first four reports provide the user with detailed information at the zip code level, specifically giving current progress within those zip codes. The focus of this research is on the fifth report, the Tactical Segmentation Market Report. This report can also provide information at multiple other levels, including the recruiting station (RS), the recruiting company (RTC), the recruiting battalion (RBN), and the brigade.

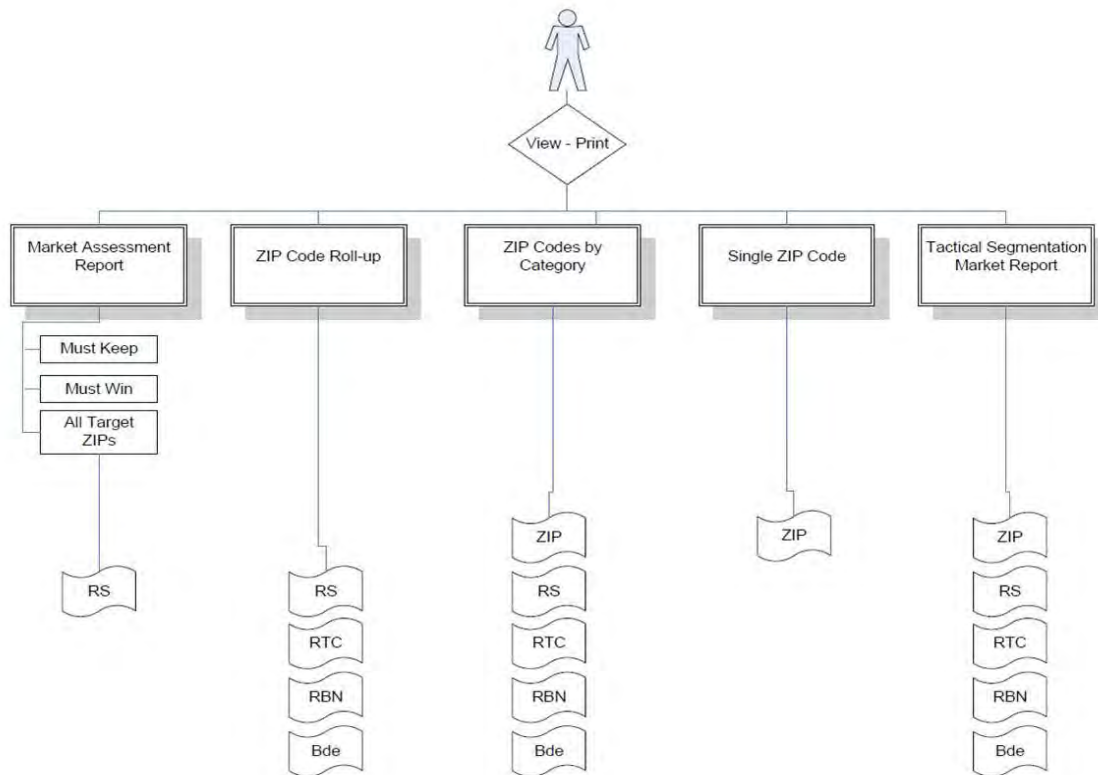



Figure 4. SAMA Reports Overview (from USAREC G2, 2012).

The results of this report are also used in the other four reports. The report partitions a unit's Area of Operation into 39 Army Tactical Segments. These were formed using Claritas Potential Rating Index for Zip Markets, New Evolution (PRIZM NE) combined with data depicting demographic differences in attitudes toward the Army (Clingan, 2007). Claritas PRIZM NE is a set of segments based on syndicated survey results and research owned by the Nielsen Company. It is a widely used customer segmentation system that is annually updated (Claritas, 2015). The Army Custom Segment Ground Counts, however, have not been updated since November 2006.



Tactical Segmentation Market Report
Army Prod As Of: 12-Aug-2009
RSD: 1A1G - JOHNSTOWN

Tactical Segment	Strategic Segment	Ts Population	Ts Population Percent	YTD Production	YTD Production Percent	Current Index	YTD Production Previous Yr	YTD Production Previous Yr Percent	YTD Production Previous Yr Index	4 Yr Wt Avg PROD	4 Yr Wt Avg PROD %	4 Yr Wt Avg PROD Index
01	01	30.32	0.25%	0	0.00%	0.00	0	0.00%	0.00	0	0.00%	0.00
02	01	139.16	1.17%	1	3.70%	3.17	0	0.00%	0.00	0.25	3.45%	2.95
03		1,121.04	9.40%	0	0.00%	0.00	0	0.00%	0.00	0	0.00%	0.00
04	02	225.86	1.89%	2	7.41%	3.91	0	0.00%	0.00	0	0.00%	0.00
05	02	163.3	1.37%	0	0.00%	0.00	0	0.00%	0.00	0	0.00%	0.00
37	10	167.28	1.40%	0	0.00%	0.00	0	0.00%	0.00	0	0.00%	0.00
38	10	271.91	2.26%	0	0.00%	0.00	0	0.00%	0.00	0	0.00%	0.00
39	10	708.1	5.94%	0	0.00%	0.00	0	0.00%	0.00	0	0.00%	0.00
NC		0	0.00%	0	0.00%	0.00	0	0.00%	0.00	0	0.00%	0.00
	NC	0	0.00%	2	7.41%	0.00	1	20.00%	0.00	1.5	20.69%	0.00
TOTAL		11,923.86	100 %	27	100 %		5	100 %		7.25	100 %	

TACTICAL SEGMENT Lists TS01-TS39 and Not Coded

STRATEGIC SEGMENT Lists which Strategic Segment the TS falls into

TS POPULATION Raw TS population data (note that "Not Coded" will equal 0)

TACTICAL SEGMENT TS proportion of the total TS population for the territory

TACTICAL SEGMENT Lists TS01-TS39 and Not Coded

YTD PRODUCTION The current production as calculated by TgtProdAchv in the Market Assessment Report

YTD PRODUCTION % The proportion of the total YTD production for that territory

CURRENT INDEX Calculated as YTD Production (%) / TS Population (%) for that segment

YTD PROD PREV YR Based on the Army Production as of <<date>>

YTD PROD PREV YR % The proportion of the YTD production for the previous year for the territory

YTD PROD PREV YR INDEX Calculated as YTD Production Previous Year (%) / TS Population (%) for that segment

4 Yr Wt Avg PROD Calculated as 4YrAvgContracts from Market Assessment Report

4 Yr Wt Avg PROD % Average production proportion of the total territory population

4 Yr Wt Avg PROD INDEX 4YrAvgContracts % / TS Population % for that segment

Figure 5. Tactical Segmentation Market Report Example
(from USAREC G2, 2012).

The Tactical Segmentation Market Report aggregates contract performance by each Army Tactical Segment. The report provides for each Tactical Segment the population size, Year-To-Date (YTD) production, previous YTD production, and a four-year weighted average production. These metrics are used by the SAMA tool to calculate the potential for a recruiting center. This methodology is detailed in Appendix A. In summary, SAMA calculates a center's best recruiting penetration rate for each Tactical Segment by finding the rate of enlistments for both the center and its company. The maximum of these two values is defined as the center's potential for that Tactical Segment. Inherently, this leads to potential as a strict measure of the highest previous performance within each tactical segment by the center or company. SAMA sets the

potential penetration rate of the best performing center to its target penetration rate for all centers with no directive for improvement. Centers performing strongly in certain tactical segments may result in unfair predictions for other centers within the company. The lower performance of some centers may be due to factors which do not apply to other centers (population density of the centers' areas, unemployment rates, ratio of recruiter force, etc.).

C. PREVIOUS EXPLORATION

In 2014, David Devin (USAREC G-2 Market Analyst) conducted an initial SAMA Methodology Validation in response to Nashville Battalion Quarterly Training Brief (QTB) in October 2013. USAREC G2 noted large differences in combined penetration rates between centers and company for the Clarksville Recruiting Company. The penetration (see Table 1) rates for each recruiting station in the company, with its associated recruiting station identification number (RSID), show the large difference between Paducah Center and the other centers in particular. Table 1 displays the penetration rate for each unit's performance in the Regular Army (RA), the Army Reserve (AR), and combined (both RA and AR).

RSID	RA	AR	Combined
5N5 Clarksville RTC	4.25	0.73	4.99
5N5C Clarksville	5.70	1.03	6.72
5N5M Paducah	2.29	0.36	2.64
5N5S Hopkinsville	4.62	0.78	5.40

Table 1. Clarksville Company Data for 2013, Company and Centers
(from Devin, 2014)

Devin's problem for analysis is as follows: "Does the SAMA method of applying the higher of two penetration rates (company or center) to determine potential result in a reasonable level of potential given this disparity between Paducah Center and Clarksville

Company?” (Devin, 2014). One initial disparity between the findings at the QTB and SAMA calculations is that SAMA only uses Active Duty data, not Army Reserve or combined data.

When conducting statistical analysis on SAMA data, the potentials must be calculated manually for a given time period since SAMA itself is a real-time calculation tool and does not store historical calculations for the interested units to compare potential between the centers and the company. Table 2 gives the SAMA potentials for the Clarksville centers and company for the RA. The Army RA Potential measure represents the SAMA calculated potential, the “Army RA 4yr_Wtd_Avg” is the four-year weighted average number of enlistments in the RA, and the “Potential Over / Under the four-year Avg” is the difference between the two.

Unit	Army RA Potential	Army RA 4yr_Wtd_Avg	Potential Over / Under the 4-year Avg
Clarksville Company	410.3	338.8	71.5
Clarksville Center	197.4	198.8	-1.4
Paducah Center	110.8	61	49.8
Hopkinsville Center	102	79	23

Table 2. SAMA Potential Calculations for Clarksville Company as of September 2014 (from Devin, 2014)

Devin’s analysis focuses on potential calculation by tactical segment for Clarksville Company and its centers, with particular interest in identifying those segments that most account for the change in actual and potential penetration rates. Additionally, he investigates the difference in segment populations (rural vs. urban). In summary, Devin concludes that the Paducah Center’s potential is achievable, but requires more resources and that additional factors influence the actual penetration rate, particularly population density and the number of recruiters assigned. These findings strongly suggest that the introduction of additional factors into a prediction model within SAMA is needed.

Devin also conducts an analysis of the current SAMA method for all centers. He defines an index in order to compare calculated potential against previous performance. The index is formulated by dividing the SAMA calculated Center Potential by the Center Four Year Weighted Average (4yr_Wtd_Avg). An index of a 1.00 signifies that the calculated potential is equal to historic past production. Similarly, a center index of 1.50 means that the calculated potential for a center is a 50% increase over its historic past production. Figure 6 outlines the SAMA potential to actual calculation using this index for 843 recruiting center.

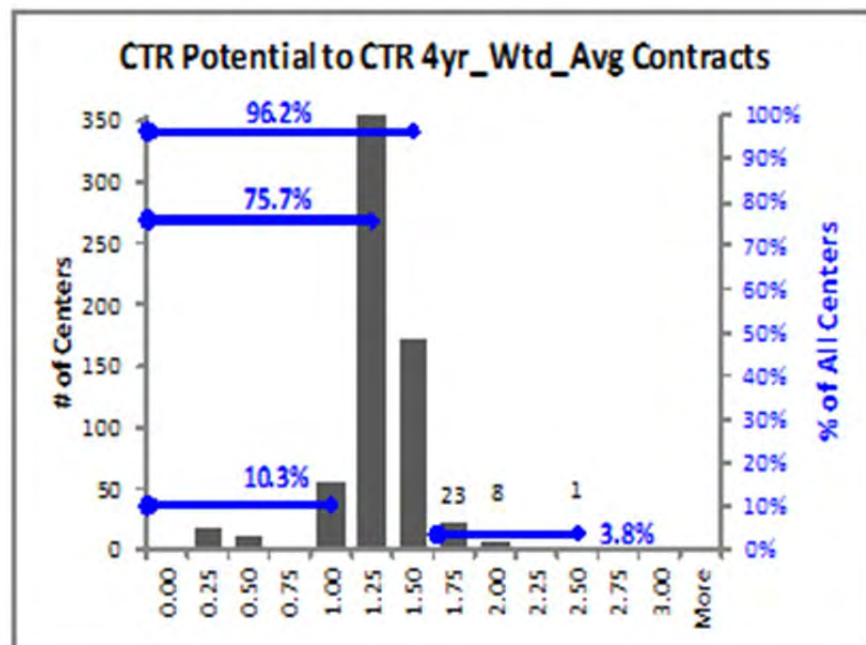


Figure 6. SAMA Potential to Actual for 843 Centers (from Devin, SAMA Methodology Validation, 2014)

The calculated potential for 96% of the centers is 25% greater than the historic past production, with 3.8% of the centers having a 50% greater calculated potential than their historic past production. These findings suggest an inflated prediction of potential. Devin investigates several other alternative approaches to potential calculation, including a USAREC level potential penetration rate application, an urbanity potential rate, and a

capping system of penetration rate. The USAREC method and the urbanity method result in a greater disparity between calculated potentials and actual performances. The capping methods decrease the disparity, but in exchange for a decrease in overall estimated potential.

In summary, Devin makes several conclusions that contribute to the background of this research. Calculated potential has limited scalability and calculated potential is best used for relative comparisons to determine where to allocate manpower and resources. Ultimately, the current calculation methods generally inflate the potential over the actual performance. To the user, this could promote frustrations or a lack of trust in a potential that is regularly out of reach.

D. OTHER RELATED WORKS

Penney, Horgen, and Borman of the U.S. Army Research Institute for the Behavioral and Social Sciences compile a technical report containing an annotated bibliography of research on Army recruiting (Penney, Horgen, & Borman, 2000). The majority of this work covers the time period of 1980 to 1999. The report identifies factors that are significant to successful recruiting of an all-volunteer force. Many different factors are investigated, including personal characteristics of recruiters, training and development, advertising support, and environmental factors. Some of these factors are shown to be very influential in a recruit's propensity to enlist. Unemployment rate, urban population, regional unemployment, educational benefits, and the number of recruiters in an area generally lead to increased enlistments. The authors use a multiple regression, pooled cross-section, time-series model to determine these findings. Though some of these factors could prove infeasible for use in current prediction models, factors extracted from data sets such as ones for unemployment and recruiter force are readily available to USAREC and could possibly be implemented in future modelling approaches provided these factors help predict an individual's propensity to enlist.

DeReu and Robbin combine applied available geo-demographic segmentation data from the Claritas Company with more traditional factors related to recruiting such as QMA and past production to study Army recruiting. Their regression model did much to

explain variability in production, and at the time, was used to assign mission (Dereu & Robbin, 1983). This mission assignment process was used in the early 1980s and differs greatly from the more basic model used today. However, integrating early geo-demographic segmentation data with the previously used factors resulted in a model with a slightly higher fidelity, though the demographic data at the time was still in the early stages of its uses. The geo-demographic segmentation data is used to partition zip codes into 34 orthogonal factors which explain 87% of the common variance of the social measures. Current segmentation data, the PRIZM NE, has consistently evolved in its reliability and is used by thousands of marketers within Fortune 500 companies (Claritas, 2015).

More recently, Williams investigates factors influencing U.S. Navy recruiting production. Williams analyzes the Noble Index, a model developed with the purpose of determining market potential of a specific geographic area. He also develops several annual and monthly models for predicting recruiting potential using both linear models and multiple regression models. Important to this research is the identification of distances from recruiting stations to Military Entrance Processing Stations (MEPS) as well as QMA as significant factors (Williams, 2014).

III. DATA COLLECTION AND PREPARATION

A. SAMA SPECIFIC ANALYSIS

Examination of current USAREC methodologies for missioning and measuring of potential as well as research on previous related studies helps give a broad idea of the type of data needed for our research. We replicate four years of SAMA results using previous enlistment contracting performance, Army Tactical Segment populations, and the number of contracts obtained within each of the Army Tactical Segments. USAREC G2 provided five years of contract performance data for all recruiting centers from 2010 to 2014, including the specific performance and population data within each tactical segment. In order to adequately evaluate current SAMA calculation methods using PRIZM NE data in place of the Army Tactical Segment data, a similar performance and population data set by the 66 PRIZM NE segments is required. From these raw data sets, a new data set is constructed to reflect SAMA output for all recruiting centers.

B. ADDITIONAL FACTORS

From other related studies such as DeReu and Robbin (1983) and Williams (2014), some factors stand out as worthy of investigation:

- Department of Defense Performance by Zip Code, 2010–2014
 - The number of zip categorized by population size as:
 - o METRO = Population over 50,000
 - o MICRO = Population between 10,000 and 50,000
 - o OTHER = Population less than 10,000
- Unemployment rate for each recruiting center and company
- Number of recruiters assigned to each center
- Responsibility Area (in square miles) for each recruiting center
- Distance from centers to nearest MEPS

- Drive time from centers to nearest MEPS
- QMA for each recruiting center

Much of this data comes in different formats which vary between zip-code level and recruiting center specifications, but is consolidated into a master file using Visual Basic for Applications (VBA) coding in a Macro Enabled Excel Workbook as shown in Appendix B. Table 3 gives an example of several records in the consolidated master file for different centers, represented by their RSID.

RSID	YEAR	MONTH	UNEMPLOYMENT RATE	QMA	Recruiters	Metro	Micro	Other
5D6F	2012	1	8.01	17591	7	17	0	0
5D6W	2012	1	8.01	21130	8	26	0	8
5D7F	2012	1	7.75	39458	10	48	0	1
5D7H	2012	1	7.75	35939	7	15	0	0
5D7K	2012	1	7.75	28406	6	32	0	0

Table 3. Sample of Consolidated Data for Additional Factors

C. SAMA CALCULATOR AND CONSOLIDATED DATASET

Prior to constructing a SAMA calculator, significant data cleaning and formatting is required. Due to the multiple realignments, openings, and closings of recruiting centers as part of USAREC's transition to "Recruiting Operations" (USAREC, 2012), much of the data is not consistent over the four-year period. Having four years of consistent performance is critical to making prediction calculations using the current SAMA method. Additionally, population and performance ground counts under Army Tactical Segments and PRIZM NE segments are unavailable for many recruiting centers that are not on the mainland of the United States. This again makes manual calculation of notional SAMA output impossible. Data for centers either not containing four years of consistent data or not containing segmented ground counts are deleted. Another key element to the SAMA potential calculation is performance data within each segment for the next higher unit, which in this case is the company. Jackson (2015), in a concurrent, related study, implements a Python script using the PANDAS add-on package which

consolidates the center data into company, battalion, and brigade data. We use her script to consolidate the center data to the company level.

A SAMA calculator is constructed in Microsoft Excel using 40 separate spreadsheets. The spreadsheets consist of the population data and performance data for the selected recruiting centers and companies for the time period 2010 to 2014. The base sheet requires input for the center RSID and company RSID. Once input, the calculator gives the following metrics for the chosen center:

- Tactical Segment Raw Potential
- Tactical Segment Ground Count Penetration Rate
- Tactical Segment Potential Penetration Rate
- PRIZM NE Segment Raw Potential
- PRIZM NE Segment Ground Count Penetration Rate
- PRIZM NE Segment Potential Penetration Rate

Additionally, this calculator supplements these outputs with additional factors for examination through cross-referencing using VBA. Finally, a looping algorithm is implemented to cycle through the calculations for each center and company, copying them to a single data file. This file consists of 750 observations (centers) with 581 columns.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS OF CURRENT SAMA CALCULATIONS

Through his use of the SAMA index (center SAMA potential/center four-year weighted average enlistments), Devin (2014) shows that SAMA over-predicts potential by more than 25% for 96% of the centers. In Figure 7, we see that SAMA calculates an average potential of over 35% over the Recruiter Year 2014 performance. Figure 9 also shows a comparison of the mean SAMA index by brigade, with the dots representing individual centers in those brigades.

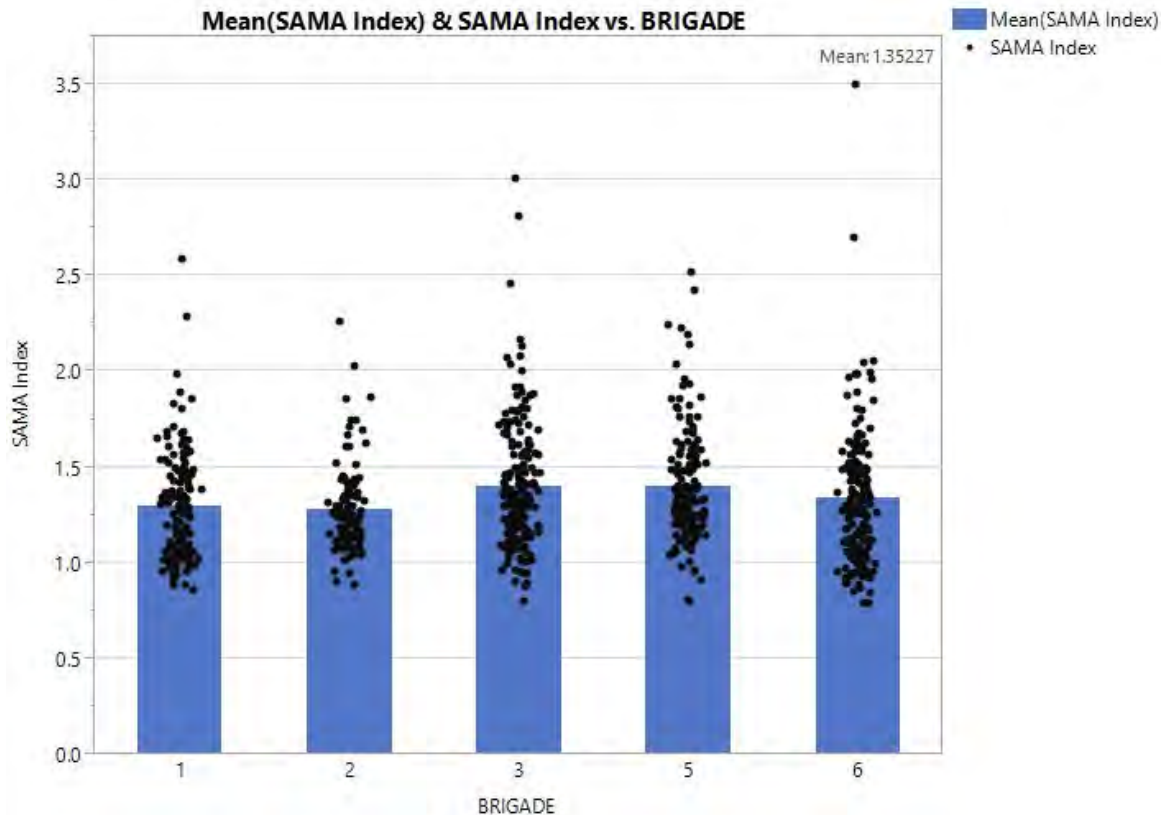


Figure 7. SAMA Index Average by Brigade including individual centers

There does not appear to be much disparity between brigades for SAMA index level, and there are only a few outlier centers, typically on the upper end where SAMA grossly over-predicts their potential (or they severely underperform in 2014).

A. CURRENT SAMA CALCULATIONS VS. 2014 ACHIEVEMENT

To conduct a more in depth look at the current SAMA potential calculation, we fit a simple linear regression in order to quantify the relationship between the calculated SAMA potential and the actual 2014 performance. The scatterplot in Figure 8 shows a strong linear relationship between the SAMA calculated potential and the contracting achievement for 2014, with an R-squared value of 0.871.

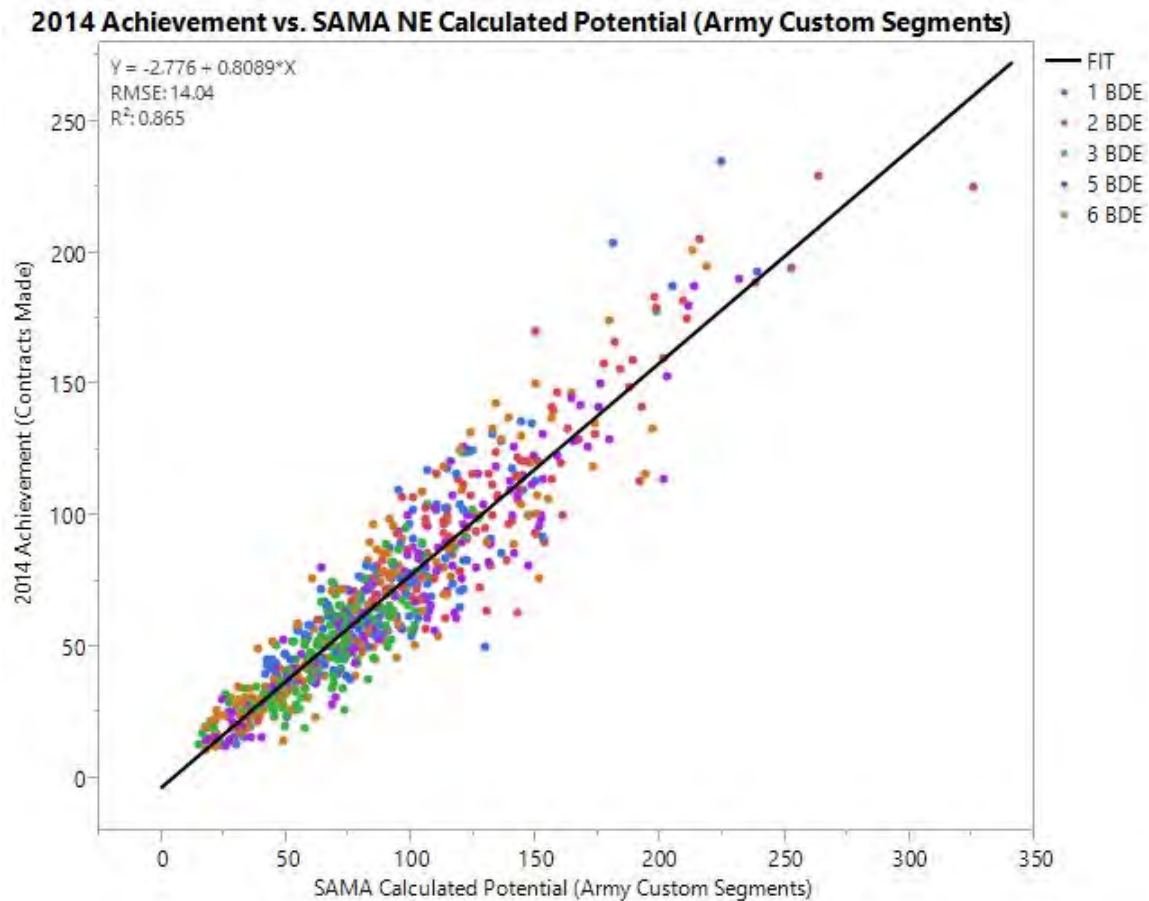


Figure 8. Simple Linear Regression of 2014 Contract Achievement by SAMA Calculated Potential (from Army Custom Segments)

The R-squared statistic measures the percentage of the variance in the response (2014 Achievement) explained by the model. We also note that the variance of the response variable is not constant; the higher performing centers' 2014 achievements are more variable and hence more difficult to predict.

B. PRIZM NE SAMA CALCULATIONS VERSUS 2014 ACHIEVEMENT

Before investigating the cause for the consistent inflation of potential based on current SAMA methodology, it is worth exploring a planned change to SAMA calculations in the near future. The current methods use the 39 Army Custom Segments for partitioning the population of the given unit and its next higher unit. It then takes the highest penetration rate of the two to set the “standard” as the potential for that segment. This translates to 39 opportunities for either “setting the standard” for a segment or having room for improvement. As of the time of this research, USAREC is committed to transitioning to the use of the PRIZM NE segments by Nielsen in lieu of the outdated Army Custom Segments (M. Stokan, personal communication, December 16, 2014). Using a similar SAMA calculations method, there will now be 66 partitions. We compare the two different methods by constructing the SAMA potential index using the PRIZM NE segments. We call this the PRIZM NE SAMA Index. Figure 9 shows a comparison of the mean SAMA PRIZM NE Index by brigade, with the dots representing individual centers within those brigades.

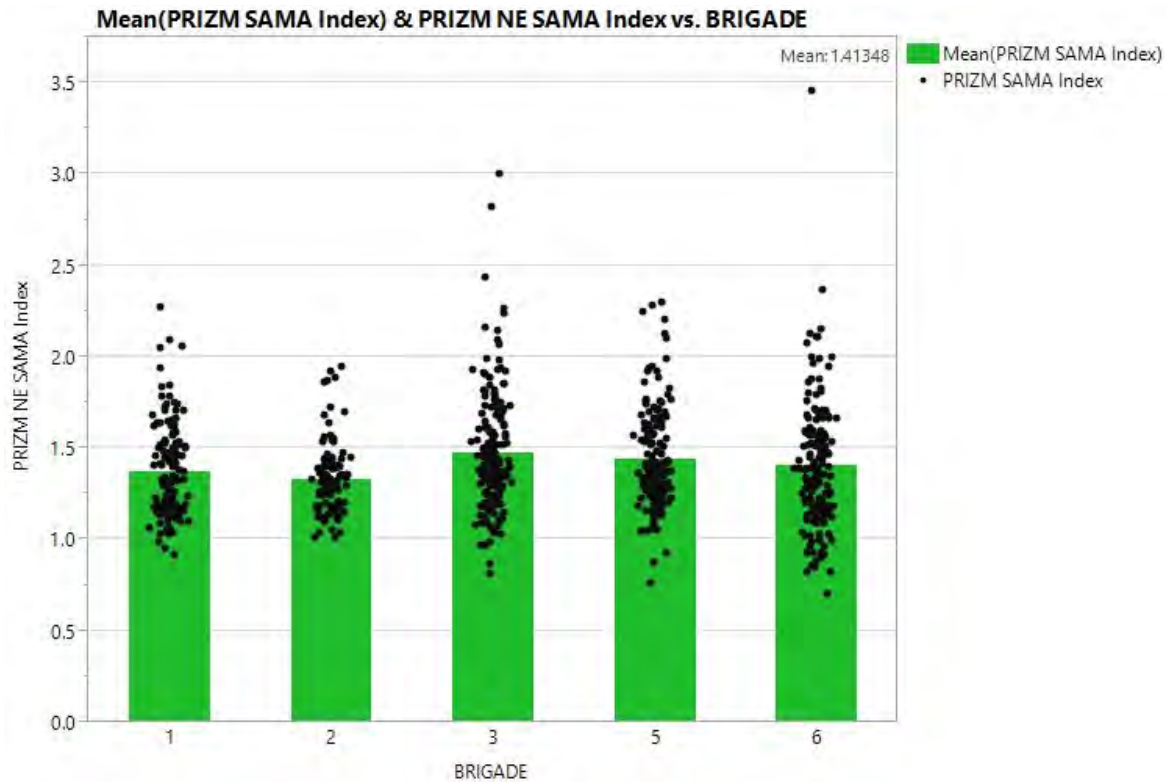


Figure 9. SAMA PRIZM Index Average by Brigade including individual centers

The graphical output in Figure 9 appears similar to the plot of current SAMA Index against production of Figure 7 with the exception of the mean. SAMA using the PRIZM NE segments calculates an average potential of over 41% over the 2014 performance, a 6% increase over the currently used model. This increase will result in even larger gaps between the performance and the levels of potential for the individual centers that are using the SAMA tool. When fitting a regression line using least squares to fit a straight line model for this data, the graphical outputs return a very similar looking fit in Figure 10 to that of Figure 8 as well as extremely similar patterns in the residual. This means that the assumption for homogeneous variance in the residuals is once again violated. However, there is a slightly improved R-squared (0.898) and adjusted R-squared (0.898), signifying marginally better prediction accuracy of the PRIZM NE.

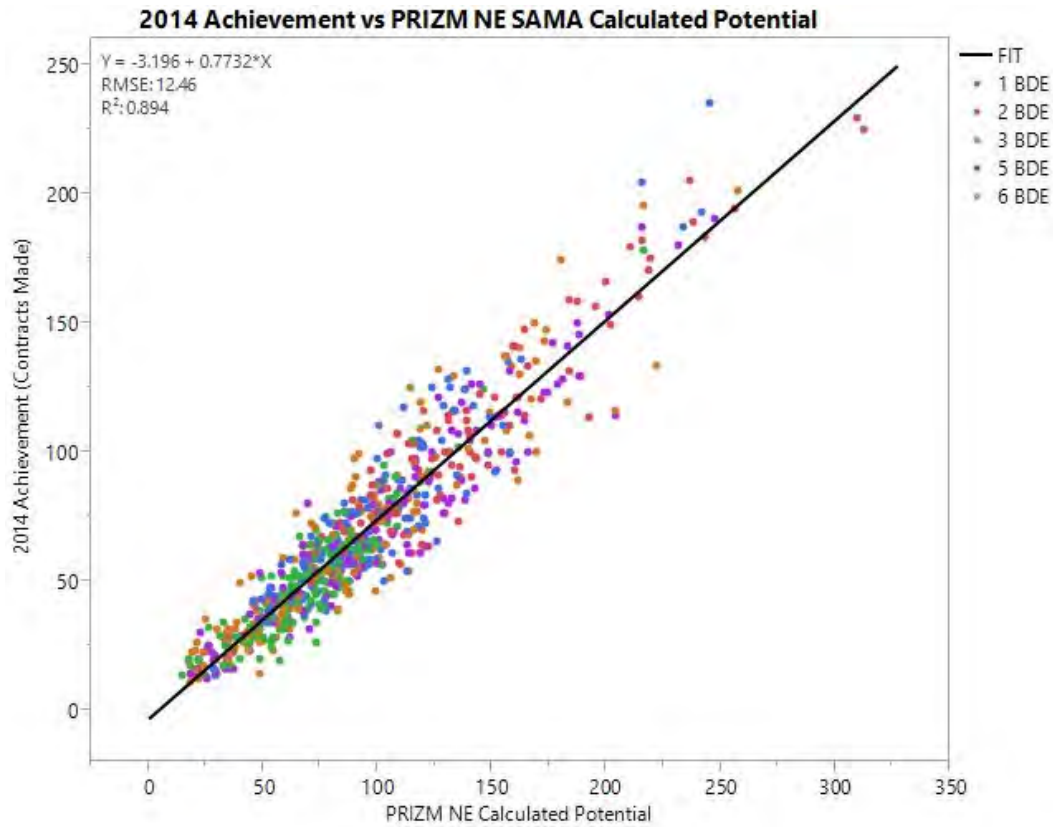


Figure 10. Simple Linear Regression of 2014 Contract Achievement by SAMA PRIZM NE Calculated Potential

Considering the very small difference between the measures of fit, there appears to be no significant benefit to switching to using the PRIZM NE data for the current SAMA calculation technique. Seeing as the Army Custom Segmentation data is outdated and is not planned for updating, the switch does not have an adverse effect either.

THIS PAGE INTENTIONALLY LEFT BLANK

V. MODELING

The current SAMA potential calculations yield close, but consistent, over-predictions of a unit's recruitment contracts. However, the model itself does not meet all assumptions to qualify for classification as an adequate model and relies only on previous performance as a factor. Previous performance factors alone do not allow flexibility for changes in market trends, command influence, or policies within the organization. Additionally, the Commanding General challenges USAREC G2 to explore planning and predictive tools for the recruiting force that do not use previous performance as a factor in calculating predictions and potential (M. Stokan, personal communication, December 16, 2014). With this consideration in mind as well as the multiple other identified factors that are likely to have influence on recruiting performance, we explore several modeling techniques and show different methods in predicting a center's recruiting potential.

A. PRIZM NE CONSOLIDATION SCORES AS SIMPLIFIED FACTORS

When exploring possible regression modeling techniques, it is common to keep a model as simple as possible in terms of number of factors while maintaining strong predictive power. We now use the PRIZM NE data instead of the outdated Army Custom Segmentation data. The 66 separate segments under PRIZM NE, if left unmodified, directly translate to 66 factors when using Least Squares or Generalized Linear Models. Looking at Figure 11, a bar chart of annual performance for contracts by segment, we see that some segments yield many fewer contracts than other segments.

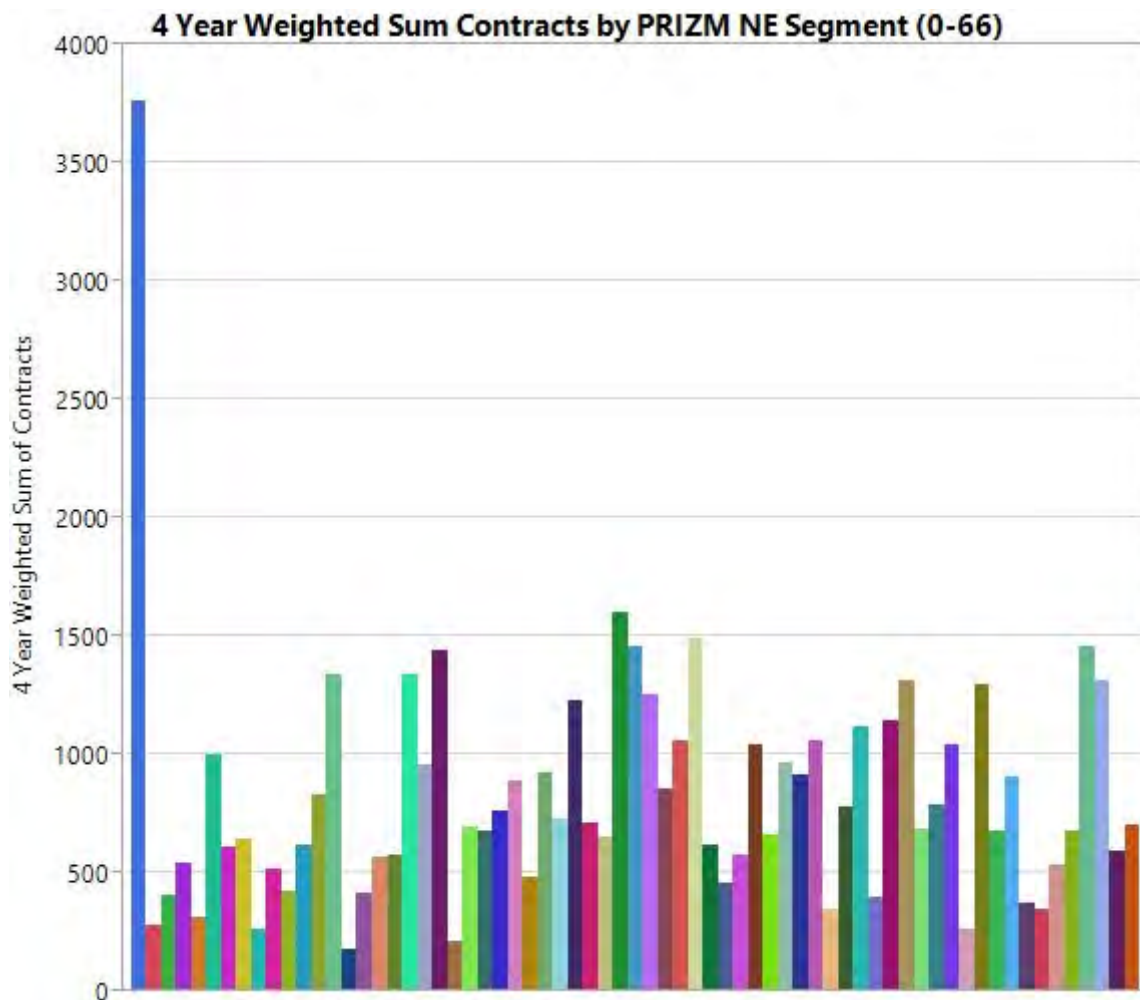


Figure 11. Four Year (2010–2014) Weighted Average of Total Contracts per PRIZM NE Segment

It is not surprising that many PRIZM NE segments within each RSID yield no recruits in any of the four years. This is expected due to the large number of population partitions native to the PRIZM NE segmentations compared to the overall number of contracts achieved by each RSID. In order to simplify the application of the segmentation data as a factor for predictive modeling purposes, we look at creating single scores that represented each RSID’s population representation by the 66 PRIZM NE segments.

1. CORE Score

USAREC G2 provides consolidated data of recruiting performance by PRIZM NE segment which further identifies 14 “core” segments out of the 66 (Baird, 2014).

Analyzing the range of recruiting performance for each core segment and identifying and isolating those high-performance segments, 14 segments are identified that produce over 35% of the recruitment contracts while only 24% of the population of the country is represented by these segments. These core segments are displayed in Table 4.

Segment	Segment Name	QMA	Contracts	% pop	% prod
13	Upward Bound	669886.6153	8135	1.98%	2.55%
18	Kids & Cul-de-Sacs	599818.3966	8294	1.77%	2.60%
20	Fast-Track Families	552454.9856	8561	1.63%	2.69%
32	New Homesteaders	629386.3684	10063	1.86%	3.16%
33	Big Sky Families	571184.7623	9051	1.69%	2.84%
34	White Picket Fences	590995.2	7576	1.75%	2.38%
36	Blue-Chip Blues	522695.8117	6707	1.55%	2.10%
37	Mayberry-ville	694944.9646	9355	2.05%	2.93%
41	Sunset City Blues	483575.5728	6515	1.43%	2.04%
45	Blue Highways	527214.7064	6419	1.56%	2.01%
50	Kid Country, USA	511345.5095	7319	1.51%	2.30%
51	Shotguns & Pickups	554067.3129	8129	1.64%	2.55%
56	Crossroads Villagers	599244.594	8011	1.77%	2.51%
64	Bedrock America	714056.3212	8346	2.11%	2.62%
Total			112481	24.30%	35.28%

Table 4. CORE PRIZM NE Segments

This suggests that the type of segments within each RSID impacts the propensity of a population to enlist. However, applying these finding in a raw form does very little to differentiate the actual performance of each RSID by PRIZM NE segment. We calculate the penetration rate of each PRIZM NE segment by dividing the percentage of the contracts from that segment by the percentage of the population in that segment. This is similar to how the penetration rate is calculated for each Army Custom Segment within the SAMA calculations. Organizing the population of each RSID by PRIZM NE segment and the penetration rates, we use the Microsoft Excel function “MMULT” to conduct matrix multiplication for each segment, divide by total population, and get a single PRIZM NE penetration score applicable to that RSID. This score does not account for any previous performance by the related RSID but rather uses the performance of that segment country-wide and weights it against the population percentage represented within the RSID by that segment. Within our data set, this score is termed “CORE Score.”

2. SOCIAL Score and LIFE Score

The Claritas company also identifies “Social Groups” and “Lifestage Groups” (Claritas, 2015). The Social Groups consist of 14 separate groups of PRIZM NE segments based on Nielsen Urbanization class and affluence. Initially, the 66 segments are placed in one of four urban categories, where they are then grouped based on affluence. Figure 12 displays these groupings.



Figure 12. PRIZM NE Social Groups (from Claritas, 2015)

Similarly, the Lifestage Groups consist of 11 separate groups of PRIZM NE segments which are based on affluence and a combination of household age and household composition (number of children). First, the 66 segments are placed in one of three Lifestage classes (Younger Years, Family Life, and Mature Years, where they are

then grouped based on affluence, household age, and the number of children in the household. Figure 13 displays these categorizations.



Figure 13. PRIZM NE Lifestage Groups (from Claritas, 2015)

Reorganizing the raw performance data by PRIZM NE segment, we can calculate the penetration rates of each Social Group and Lifestage Group as well as the populations represented by each Social Group and Lifestage Group. Using similar calculation and

matrix multiplication techniques as used in calculating the CORE Scores, scores termed “SOCIAL Score” and “LIFE Score” are calculated for each RSID. This allows us to simplify the data by consolidating the PRIZM NE factors into the CORE, SOCIAL, and LIFE scores. This allows for easier examination of data as well as checks for significance, correlation, and the regression model, particularly when using weighted least squares.

B. EXAMINATION OF DATA

In preparation for formulating models, the data set is first examined by plotting the response “y” against each of the potential “x’s.” From these one to one plots, we get a general idea of the predictors that are significantly related to the response. Appendix C gives plots and summary statistics used during the examination of the data prior to formulating the models. The response (2014 Contract Achievement) is plotted against the following factors individually:

- Weighted four-year average enlistments
- Weighted four-year average enlistment for all services
- QMA (Qualified Military Available, aged 17–24 year)
- Recruiters (number of recruiters assigned to recruiting center)
- Unemployment Rate
- Driving distance in miles from center to nearest military processing station
- Driving time from center to nearest military processing station
- Square mileage of the center’s area of responsibility
- Score based on representation of high performing segments within a center (CORE Score)
- Score based on representation of high performing social groups within a center (SOCIAL Score)

- Score based on representation of high performing lifestyle groups within a center (LIFE Score)

Based on these plots (Figure 21, Appendix C), we eliminate square mileage as a factor and determine that the four-year weighted average variable and the four-year DOD average have a strong linear relationship with the response. All of the other factors are correlated with the response, but do not show as strong a linear relationship as the previous performance factors. The two most strongly related factors are purely based on previous performance, which we would like to try to exclude in several models in an attempt to formulate a model that does not depend on individual previous performance as a factor.

Following the individual plots, scatter plots for all pairs of predictors are displayed as a scatterplot matrix (Figure 22, Appendix C). Table 4 in Appendix C is the correlation matrix for the predictor variables. The correlation coefficient between two variables shows the measure of linear association, which is a value between -1 and 1. The scatter plot and correlation matrix shown in Appendix C help identify the strong correlation between MEPS distance and MEPS driving time as well as between LIFE score and CORE score. These factors as pairs should not be included in the modelling efforts together in order to reduce redundancy and the number of unnecessary terms.

C. MULTIPLE LINEAR REGRESSION MODEL WITH PREVIOUS PERFORMANCE

Regression is one of the most popular applied statistical techniques that is commonly used for prediction, particularly for forecasting a response (dependent variable) based on one or more factors which are independent or predictor variables (Klimberg & McCullough, 2013). Regression can be used to model both linear and nonlinear relationships between the response and the predictor variables with variables with both linear and non-linear relationships. A multiple linear regression equation takes the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

In this equation, y signifies the response variable, the β 's are the parameters, x 's are the predictor variables, and ε is the associated error. We note that the y 's and the x 's may be transformed versions of the original variables. Transformations of the original variables allow for a rich class of regression models that can include non-linear relationships. The method of least squares is used to estimate the coefficients. For the predictions to be accurate and for the model to be considered adequate, the model must pass the following assumptions:

- Constant variance of the errors
- Normality of the error distribution
- Statistical independence of the errors—little or no correlation
- Linear relationship between the independent and dependent variables

1. Formulating and Fitting the Model

Looking at the individual plots from examining the data as well as the individual relationships between the predictors and the response, we can see that generally linear relationships exist between the response and the predictor variables, making the addition of curvilinear terms appear unnecessary. This is checked after the model is fit and is discussed in the next section. Many approaches exist for variable selection when conducting a linear regression. The one used during the study is a stepwise regression facilitated through the JMP 11 PRO platform. Prior to fitting the regression, the 750 observations are partitioned into training, validation, and test sets (450, 150, 150 observations). The validation set is used in order to detect and avoid over-fitting, which becomes obvious when there are large discrepancies between the goodness of fits in the training and the validation sets with the same model. The test set is held apart from all model fitting efforts. It is used to obtain an unbiased estimate of prediction error. In JMP, a separate column is easily made to facilitate this partition.

When using stepwise regression, several model selection approaches are available. In this situation, the minimum Akaike Information Criterion (AIC) is chosen as the stopping rule rather than the Bayesian Information Criterion (BIC). BIC regularly

imposes a greater penalty for larger numbers of terms in a model, but AIC was chosen in order to possibly look at a larger number of terms and remove terms individually if deemed necessary. The following variables are selected thru stepwise regression using minimum AIC as a stopping rule:

- Brigade (Categorical)
- Four-Year Weighted Average Contract Performance
- CORE Score
- SOCIAL Score

The R-squared of 0.878 is high and an improvement on the current implementation of SAMA as a predictive model. Figure 14 shows the resulting plot of the actual y values against their predicted values (using the fitted regression model) along with several of the fit statistics.

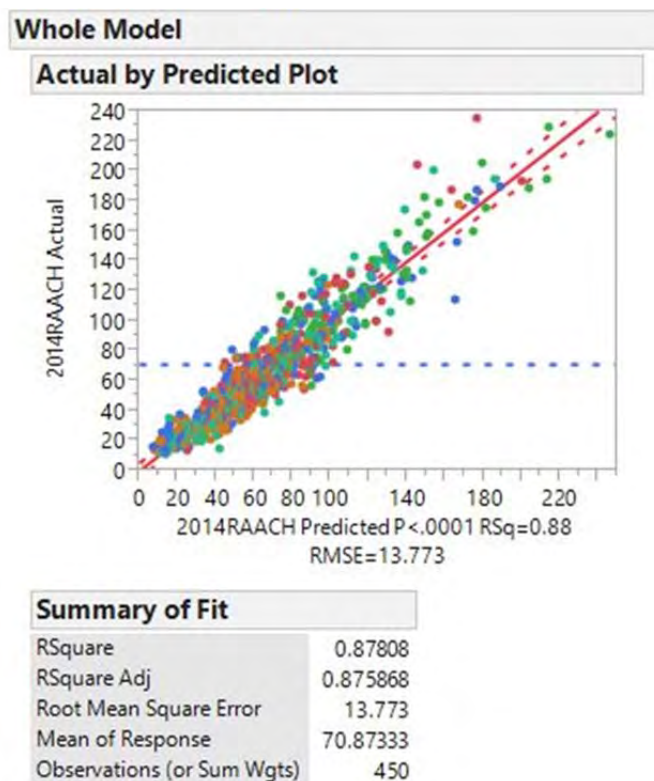


Figure 14. Initial Linear Regression Model Actual by Predicted Plot and Summary of Fit

Next, we check the fit of the model and ensure that all of the assumptions for an adequate model are met.

2. Checking Assumptions and Validation

In order to check the assumptions previously mentioned, we inspect residual plots, the Variance Inflation Factors (VIFs), the Durbin-Watson statistic, and the Cook's Distance plot for influential observations.

The first assumption checked is for constant variance in the error terms through the plot of the residuals against the predicted values. In order to meet the assumption, there should be no obvious trend or pattern present in the plot. The plot (not reproduced here) from this fit shows a distinct cone shaped pattern, with a trend of increased variance as the predictions increase. This cone shaped pattern of increasing variance is also evident in Figure 14. A common method for addressing non-constant variance is to transform the response. Transformations of the response often address non-normal residuals as well. Within the JMP interface is the option to conduct a Box-Cox transformation, which helps suggest the best power transformation of the response in order to stabilize the variance. Figure 15 shows the Box-Cox transformation plot where λ is power and a small residual sum of squares (SSE) indicates a better choice of λ .

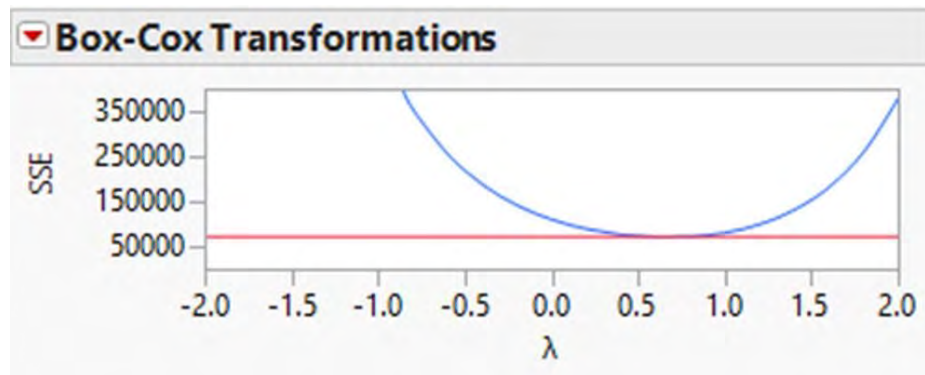


Figure 15. Box-Cox Transformation Plot for Linear Regression Model

A good transformation in this situation is a power transformation with $\lambda = 0.6$ using the following equation:

$$\frac{y^\lambda - 1}{\lambda}$$

After transforming the response variable using this power transformation and fitting the regression again using the same predictors, a linear regression model is produced with a similarly high R-squared and adjusted R-squared value, as shown in Figure 16.

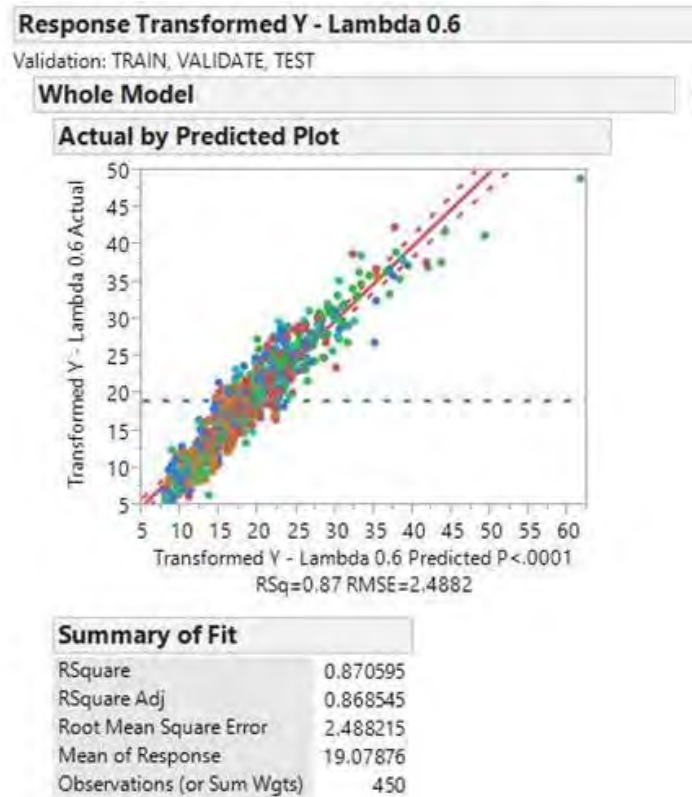


Figure 16. Linear Regression Model with Power Transformed Response

When plotting the residuals against the predicted values as shown in Figure 17, there are no longer any strong signs of non-constant variance. A small increase in variance is detected toward the right of the plot, but not enough to conclude that the assumption is violated.

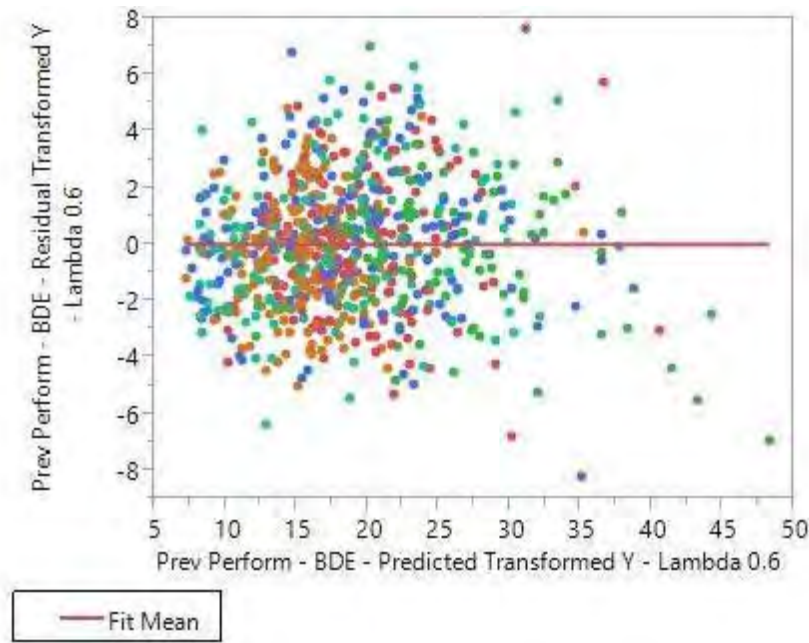


Figure 17. Residual Values against the Predicted Values plot,
Linear Regression Model

In testing the assumption for normality of the error distribution, the residuals plot accompanied by the Normal Quantile-Quantile (QQ) plot is used. A general straight line without an obvious curve leads to visually passing the assumption. Often, the Shapiro-Wilks test can also accompany a validation of this assumption. The p-value for the Shapiro-Wilks test in this situation is 0.3337, leading us to conclude that the error distribution is normal. This assumption is further confirmed through the plot and the shape of the accompanying histogram which is shown in Figure 18.

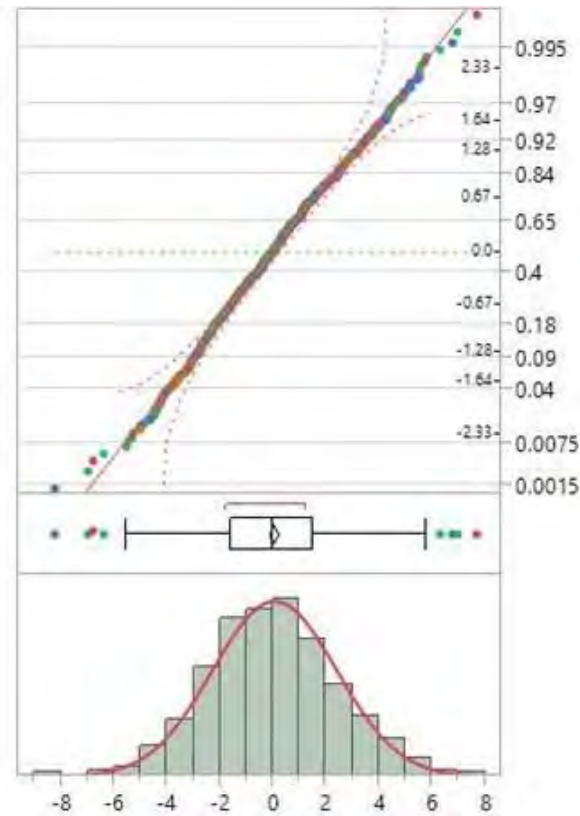


Figure 18. Residuals plot (with accompanying Normal QQ line plot),
Linear Model

Plots of residuals against each predictor show no apparent pattern confirming that even after transforming y , no curvilinear terms of the predictors are needed in the model. Further, each factor in the model has an accompanying VIF, which is displayed in the parameter estimates table in JMP. Any VIF between 5 and 10 indicates a possibility of multicollinearity, while a VIF over 10 indicates that multicollinearity is a problem and that variable should be removed. There are no VIFs higher than 10 with this fit, though CORE score and SOCIAL score each had VIFs of 8.1 and 8.4, respectively.

In order to detect possible autocorrelation, the Durbin-Watson test statistic is used. Initially, the model is suffering from autocorrelation fails the Durbin-Watson test. This is due to the data organization by brigade. This suggests that the brigade to which a center belongs contains important information for predicting; information which is not contained in other variables. We do not add a variable accounting for brigade in this model, but do add one to the model in the next section. Finally, the Cook's Distance

values are computed and plotted. These values are used to detect overly influential points. If an observation has a Cook's Distance of over 0.5, then it is deemed influential and is a candidate for removal. During the primary analysis and modeling attempts, a single influential point is found. Because we suspect that this observation represents aggregated results for two or more contracts, we remove it and refit the regression. Once removed and refit, there are no longer any influential observations present while the model once again meets all assumptions. The R-squared value for the training set is 0.8739, while the R-squared value for the validation set is 0.889, letting us safely assume that the model is not over-fit. When we transform the predicted values back to the same scale as the actual values, the R-squared value for the training set is 0.869 and the R-squared value for the validation set is 0.89 (Figure 23, Appendix D).

This model provides a high R-squared adjusted value with a low number of factors while meeting all of the model assumptions and not over-fitting. However, the four-year weighted average factor is included which directly incorporates previous performance into the model. Next, a similar modeling technique is used without using previous performance as a factor.

D. MULTIPLE LINEAR REGRESSION WITHOUT PREVIOUS PERFORMANCE

All previously introduced factors are brought into the baseline stepwise linear regression with the exception of the four-year weighted average contracts and the four-year weighted average of contracts in all of the Department of Defense. The following variables are selected through stepwise regression using minimum AIC as a stopping rule:

- Brigade (Categorical)
- QMA
- Recruiters
- Unemployment Rate
- CORE Score
- SOCIAL Score

- Driving time from center to nearest military processing station

The resulting R-squared value is 0.654 and the R-squared adjusted value is 0.646. However, the model once again fails the assumption for constant variance in the residuals against the predicted values plot. Looking at the plot for the suggested Box-Cox transformation, the ideal lambda is close to zero, making a log transformation of the response viable. Following the transformation of the response and implementing the regression using the same factors, the R-squared and R-squared adjusted values actually increase to 0.67 and 0.66, respectively. When we transform the predicted values back to the same scale as the actual values, the R-squared value for the training set is 0.582 and the R-squared value for the validation set is 0.652 (Figure 24, Appendix D). The model also passes all necessary assumptions and cross-validates well between the test and validation sets. We now try one more modeling technique to compare with multiple linear regression.

E. NEURAL NETWORK MODEL

Another approach that warrants investigation is modeling using neural networks. A neural network is made up of artificial neurons, which are often called nodes. Three types of neurons exist within a neural network: input neurons, hidden neurons, and output neurons (Yu-Wei, 2015). The strengths of connection between these neurons are called weights. The hidden nodes serve as nonlinear functions of the original inputs, and the functions applied at these nodes are the activation functions (SAS Institute, 2013). Figure 19 displays a basic artificial neural network structure.



Figure 19. Diagram of a Neural Network Architecture (from Yu-Wei, 2015)

The implementation of neural networks comes with several advantages and disadvantages. Some benefits of neural networks include their ability to detect nonlinear relationships between dependent and independent variables, giving greater flexibility to the model. Additionally, neural networks are nonparametric models, leading to the ability to eliminate errors in the estimation of parameters (Yu-Wei, 2015). The primary disadvantages include the inherent computational costs often required in fitting neural networks and the “black box” nature of the fitted model. Neural networks also tend to converge to a local minimum rather than a global minimum. Most importantly, neural networks tend to over-fit models, particularly if many hidden layers and nodes are used. The JMP 11 PRO neural networks platform also gives the option of “boosting” neural networks. With this option, a sequence of neural networks is fit and a weighted sum of these is used for prediction.

1. Formulating and Fitting the Model

The two primary decisions to make concerning hidden layers are the number of hidden layers to have within the neural network and the number of neurons there will be within each of the hidden layers. Increasing the number of layers and the number of nodes within each layer adds greater flexibility to the model but can also lead to over-

fitting the data. Using training, validation, and test set partitions allows us to prevent over-fitting our model and find the ideal level of complexity. The JMP 11 Pro neural network platform allows the implementation of two hidden layers and an unlimited number of nodes within each layer. However, increasing the number of nodes significantly increases the computation time for JMP on a standard, modern desktop computer. We use the same independent variables as identified following our initial analysis of the data and factors without using the two direct previous performance factors.

Figure 20 shows the JMP 11 PRO neural network model launch window. The “Hidden Layer Structure” dialog window allows us to input the desired number of nodes within each layer and activation function. We use only the hyperbolic tangent (TanH) activation function, which is a sigmoid function that transforms values to be between -1 and 1 and is a centered and scaled function similar to the logistic function (SAS Institute, 2013). In the “Boosting” window, we apply the number of boosted models as well as the learning rate. Increasing the number of boosted models tends to increase the precision of the model, but can also lead to over-fitting with the higher number of boosted models. When boosting, the model will fit the first network, take the residuals, and reweight observations according to the residuals (Friedman, Hastie, & Tibshirani, 2009). The process continues until the designated numbers of models are fit or until the addition of another model fails to improve the validation statistic. Learning rates (a number between zero and one) closer to zero have a lower tendency in over-fitting the data but do not converge as quickly on the final model. We choose a very small learning rate with the primary goal of avoiding over-fitting.

Neural

Validation Column: TRAIN, VALIDATE, TEST

Model Launch

Hidden Layer Structure

Number of nodes of each activation type

Activation Sigmoid Identity Radial

Layer	TanH	Linear	Gaussian
First	30	0	0
Second	10	0	0

Second layer is closer to X's in two layer models.

Boosting

Fit an additive sequence of models scaled by the learning rate.

Number of Models

13

Learning Rate

0.001

Fitting Options

☒ Transform Covariates

☒ Robust Fit

Penalty Method

Squared

▼

Number of Tours

1

Figure 20. JMP 11 PRO Neural Network Launch Window

Finally, JMP 11 PRO offers several fitting options. The transforming covariates option transforms all continuous variables to near normality which helps remove the negative influence of outliers. Selecting the robust fit option also helps minimize the influence of response outliers. Four penalty options are available (squared, absolute, weight decay, and no penalty). We choose the squared method in this case due to the relatively low number of predictor variables as well as our belief that our chosen predictor variables contribute to the predictive ability of the model. The last fitting option is the number of tours, which specifies the number of times to restart the fitting process. Each run initiates with random starting points for the parameter estimates, and the run with the best validation statistic is selected for the resulting model. A high number of

tours tend to result in a better result at the cost of computation time, so we choose five tours as a standard for all modelling attempts.

2. Model Selection

Adding to the number of hidden layers, the number of nodes per layer, and the number of boosted models generally adds to the level of precision of the model at a cost of computation time as well as an increased risk in over-fitting. When training the model, we test different ranges of values for these numbers within reasonable computation time (five minutes or less). We use the same partition of the data for training, validation, and test sets as previously used. Like with multiple linear regression, the R-squared statistic measures the percentage of the variance in the response (2014 Achievement) explained by the model. The root-mean-square error (RMSE) measurement for each model shows the sample standard deviation of the differences between the observed values and the predicted values, with a lower number being better. We create a table (Table 5) for the different models and compare the R-squared values and the RMSE values for the training and validation set for each model. The resulting R-squared and RMSE for the validation set for model 13 is the best out of the 15 different modelling attempts.

MODEL	LAYER 1	LAYER 2	LEARN RATE	# BOOSTS	TRAIN R2	TRAIN RMSE	VAL R2	VAL RMSE
1	10	0	0.01	10	0.68	24.72	0.688	21.21
2	10	0	0.01	20	0.69889	24.074	0.693	20.446
3	10	0	0.01	30	0.70314	24.58	0.687	20.5746
4	30	0	0.01	10	0.6786	25.035	0.685	20.95
5	30	0	0.01	20	0.731	23.32	0.693	20.46
6	30	0	0.01	30	0.74	23.18	0.687	20.51
7	10	0	0.001	10	0.721	23.998	0.679	20.97
8	10	0	0.001	20	0.695	24.51	0.692	20.55
9	10	0	0.001	30	0.66	25.48	0.582	20.99
10	30	0	0.001	10	0.6764	24.912	0.6856	20.745
11	30	0	0.001	20	0.702	24.73	0.693	20.55
12	30	0	0.001	30	0.6874	24.76	0.691	20.885
13	30	10	0.001	10	0.6911	24.6	0.6945	20.41
14	30	10	0.001	20	0.704	24.224	0.682	20.54
15	30	10	0.001	30	0.695	24.257	0.69	20.6

Table 5. Neural Network Model Results

This finding narrows our search range for the number of different boosted models to use in order to find the best fitted model, since the changing of this parameter has the most impact on the model compared to the number of layers and nodes. We can then fit more models with thirty nodes in the first hidden layer, ten nodes in the second layer, and with a learning rate of 0.001 while adjusting the number of boosted models around ten. Table 6 shows the results for these neural network models along with the computation time in JMP 11 for each run using a standard desktop computer with i7 processor and 16GB. The model with 7 boosts has a high R-squared value (0.6998) for the validation set along with the lowest RMSE of 20.02, leading to our selection as this model for representation as our neural network model.

MODEL	# BOOSTS	TRAIN R2	TRAIN RMSE	VAL R2	VAL RMSE	TIME
1	5	0.67752	24.74	0.6943	20.65	8s
2	6	0.65485	25.726	0.7024	21.13	12s
3	7	0.6884	24.69	0.6998	20.02	14s
4	8	0.71	24.323	0.691	20.72	16s
5	9	0.645	25.58	0.7	20.79	18s
6	10	0.706	24.27	0.689	20.67	21s
7	11	0.696	24.446	0.695	20.27	23s
8	12	0.74	23.25	0.691	20.4	24s
9	13	0.67545	24.875	0.699	20.32	26s
10	14	0.6768	24.95	0.697	20.13	29s
11	15	0.612	26.77	0.692	21.37	30s

Table 6. Neural Network Model Narrowed Results

THIS PAGE INTENTIONALLY LEFT BLANK

VI. MODEL COMPARISON AND CONCLUSION

A. MODEL COMPARISON

Following the initial data analysis and cleaning as discussed in Chapter V, we fit three models to the data set in order to provide a means of predicting an annual recruiting performance for any recruiting center given a set of independent factors. Each of these models can possibly replace existing potential calculations in SAMA or supplement the information currently provided by the SAMA platform. The first and second models implement multiple linear regressions using least squares. The first model uses two very influential factors governed by previous performance; the center four-year weighted average of enlistments and the DOD four-year weighted average of enlistments for the center's area. These factors do provide stronger immediate predictive power to the model, but tend to be inflexible as they do not allow for changes in market trends, command influence, or policies within the organization. The second model does not use these previous performance factors, depending on the factors of brigade assigned, QMA, number of recruiters assigned, the unemployment rate, the formulated CORE Score and SOCIAL Score, and the MEPS driving time. The predictive power, represented by the R-squared value for the model, is not as high as that of the first model, but the factors are generally independent of previous performances or policies, allowing for more reasonable measures of potential, particularly when changes in tactics, organizational structures, and leadership are involved. The final model, the neural network model, also does not include the previous performance factors and can provide strong predictive power, but at the cost of potential over-fitting when applied to an independent test set.

When directly comparing the three models, JMP 11 PRO has a model comparison platform that shows select comparative statistics of each model side by side. We compare the R-squared value and the RMSE for each model on the training set and the test set, and we consider if all modeling assumptions are met and if previous performance factors are used. From Table 7, we see that the first model has a higher R-squared value for both the training set and the test set while meeting the modeling assumptions. However, this model does require the use of previous performance factors, while the other two do not.

The second model meets all assumptions while not using previous performance factors. It has a considerably lower R-squared value for the training set but does well on the test set with a R-squared value of 0.633. The third model does not have any distinct advantages over the second model except that it performs much better on the training set. However, on the test set, the neural network model performs slightly worse in both categories.

MODEL	TYPE	TRANSFORM	TRAIN R2	TRAIN RMSE	TEST R2	TEST RMSE	ASSUMPTIONS MET	PREVIOUS PERFORMANCE
1	Fit Least Squares	Power Transformed (.6)	0.869	14.15	0.885	12.61	Yes	Yes
2	Fit Least Squares	Log	0.582	25.299	0.633	22.56	Yes	No
3	Neural Network	None	0.675	24.9	0.62	22.2	N/A	No

Table 7. Model Comparison

In deciding between the first two models, the two main differences are the use of previous performance factors and the R-squared values. One other consideration is the transformation used for each model. Though both use a transformation of the response, a log transformation is simpler to conduct, understand, and possibly implement in a user system. The ability to predict with a reasonable R-squared value without the use of previous performance factors is also very favorable in selecting the model.

The importance of using previous performance as a factor directs which model to use. The previous performance factor is a strong predictor but inflexible to changes in market trend as well as changes in command structure and organizational policies. However, for near term predictions such as the next year, we would chose the first model given its predictive power as well as the lack of likelihood in drastic market and organization changes within a single year. However, if predicting beyond a year or following major organizational or environmental changes, the second model should be considered.

B. CONCLUSION

Current SAMA calculations are nearly always inflated. This is because each calculation for each partition of a center's performance takes the higher of two numbers between the center and the company and sets that as a standard. If high performance outliers produce over the calculated potential, the vast majority of units will not be able to do the same, and often will produce far under the potential. This can lead to a lack of faith and motivation in the system, particularly if the measurements are used in Quarterly Training Briefs, which they currently are. The application of factors using the PRIZM NE segmentation scores in modeling techniques such as multiple linear regression and modeling using neural networks helps increase the predictive power of the models. A model fit using least squares regression provides good predictive power without a trend of gross over-predicting. The additional factors, particularly the PRIZM NE segmentation scores, can eliminate the requirement for direct previous performance factors if desired. USAREC planners can use these models in predicting the performance for the centers within the command. Center and company level leaders and planners can use the models to predict performances even with changing environmental and operational factors, such as unemployment rates, population increases, or an increase or decrease in recruiters assigned.

C. FUTURE WORK

The recommended model from this research provides an annual level prediction for all enlistments for a recruiting center, and takes segmentation data from the entire center's area of operation. Future work could diversify the model into the separate recruiting categories that recruiters receive as part of their mission. These categories include high school graduates and current high school seniors along with classifications of these two categories based on military entrance exam test scores. Additional factors that help predict performance under these categories could include high school graduation rates, regional test scores, and college attendance rates.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. SAMA POTENTIAL CALCULATIONS

All calculations are completed internally to the SAMA system prior to the display of results. The user inputs what level of report is desired (Center, Company, Battalion, or Brigade), the calculations are made, and the results are displayed up to date, which effectively makes SAMA a real time calculation tool. The following explanation of the calculations is taken from the SAMA Reports Users Guide, Version 2.

A. SEGMENTATION AND FOUR-YEAR WEIGHTED AVERAGE

For the level of unit chosen, the population (aged 16–24) is split into the Army Tactical Segments. The calculations use a 40–30–20–10 weighted calculation system (4Yr_Wtd_Avg):

$$\text{YEAR 4} = \text{CURRENT YEAR}-4 (\text{YR4}*.10)$$

$$\text{YEAR 3} = \text{CURRENT YEAR}-3 (\text{YR3}*.20)$$

$$\text{YEAR 2} = \text{CURRENT YEAR}-2 (\text{YR2}*.30)$$

$$\text{YEAR 1} = \text{CURRENT YEAR}-1 (\text{YR1}*.40)$$

B. PENETRATION RATES

The best potential penetration rate is calculated for each Tactical Segment for the unit of interest. The penetration rate gives the percentage of the population that has successfully enlisted into Active Duty.

Step 1: The four-year weighted average is calculated for each tactical segment -
$$4Yr_Wtd_Avg = (YEAR\ 1\ Enlistments * 0.4) + (YEAR\ 2\ Enlistments * 0.3) + (YEAR\ 3\ Enlistments * 0.2) + (YEAR\ 4\ Enlistments * 0.1)$$

Step 2: The penetration rate (Pen Rate) for each tactical segment is calculated by dividing the four-year weighted average of each tactical segment by the applicable population size of that tactical segment -

$$Pen\ Rate = 4Yr_Wtd_Avg / Population\ Size$$

Step 3: The maximum penetration rate between the recruiting center ($PenRate_{RS}$) and the recruiting company ($PenRate_{RTC}$) equates to the Potential Penetration Rate by Tactical Segment for a center ($PotentialPenRate_{TS}$) -

$$PotentialPenRate_{TS} = \max(PenRate_{RS}, PenRate_{RTC})$$

RSID	TS1 Pen Rate	TS2 Pen Rate	...	TS39 Pen Rate
1A1	0.002210824	0.001498641	...	0.001984435
1A1D	0.00461936	0.00069909	...	0.00139826
Max	0.004619364	0.001498641	...	0.001984435

APPENDIX B. SAMPLE VBA CONSOLIDATION CODE

This VBA code is used to consolidate the unemployment rates of each zip code to its respective recruiting center and company. Similar coding techniques are used to consolidate QMAs, Urbanicity classifications, Department of Defense performances, and number of recruiters per center.

```
Sub UnemployMacro()  
  
    Dim i As Double  
    Dim j As Double  
    Dim m As Double  
    Dim n As Double  
    Dim x As Double  
    Dim y As Double  
    Dim z As Double  
    Dim numb As Double  
    Dim numc As Double  
    Dim numd As Double  
    Dim Rate As Double  
    Dim RSID As String  
  
    n = Sheet1.Range("A2", Sheet1.Range("A2").End(xlDown)).Count 'counts the number on entries on the sheet  
    z = Sheet2.Range("B3", Sheet2.Range("B3").End(xlDown)).Count 'counts the number on companies on the sheet  
  
    For i = 1 To n  
  
        numb = i + 1  
  
        If Sheet1.Range("D" & numb) <> "NA" Then  
  
            RSID = Sheet1.Range("A" & numb).Value  
            RSID = Left(RSID, 3)  
            x = Sheet1.Range("B" & numb).Value  
            y = Sheet1.Range("C" & numb).Value  
  
            For j = 1 To z  
  
                numc = j + 2  
  
                If Sheet2.Range("B" & numc).Value = RSID Then  
  
                    numd = 3  
  
                    Do While numd <= 48  
  
                        If Sheet2.Cells(1, numd) = x And Sheet2.Cells(2, numd) = y Then  
  
                            Rate = Sheet2.Cells(numc, numd)  
  
                            Sheet1.Range("D" & numb).Value = Rate  
  
                            Exit Do  
  
                        End If  
  
                        numd = numd + 1  
                    Loop  
                End If  
            Next j  
        End If  
    Next i  
End Sub
```

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. DATA EXAMINATION

Prior to formulation of models, the data is first examined. First, a plot of the response “y” for each of the potential predictors is conducted in order to investigate potential relationships with the response. The response is the number of enlistments for a center for 2014. Figures 21, 22, 23, and 24 displays these plots.



Figure 21. Plots of the response (2014 Number of Contracts) QMA, Recruiters, and Unemployment Rate

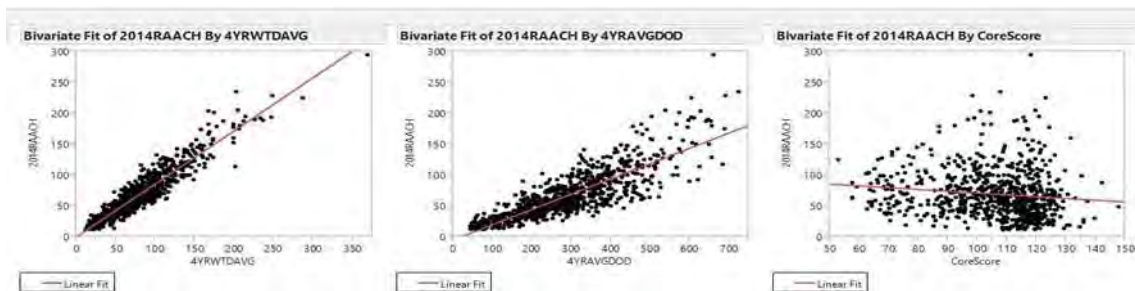


Figure 22. Plots of the response (2014 Number of Contracts) against weighted four-year average enlistments, weighted four-year average enlistments for all services, and CORE Score

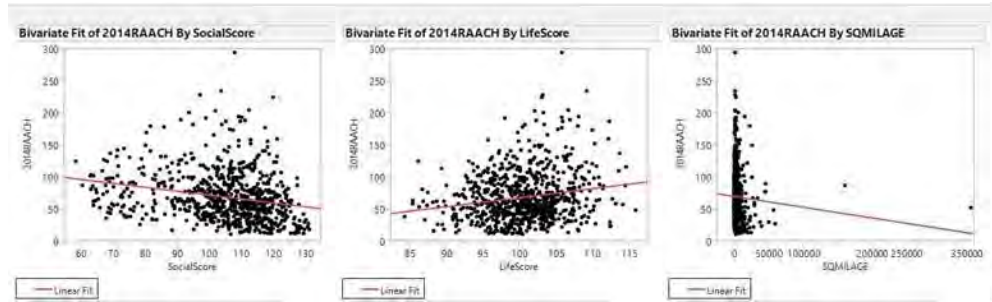


Figure 23. Plots of the response (2014 Number of Contracts) against SOCIAL Score, LIFE Score, and square mileage of center area of responsibility

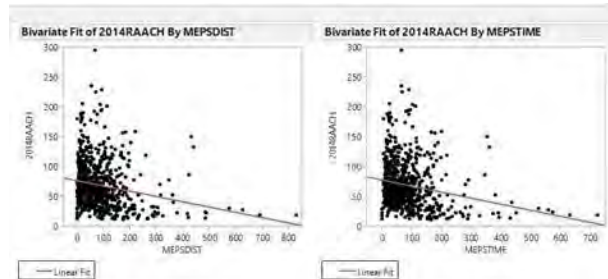


Figure 24. Plots of the response (2014 Number of Contracts) against distance (miles) from center to nearest MEPS, and driving time from center to nearest MEPS

From these plots, we eliminate square mileage as a factor. It is also clear that the four-year weighted average variable as well as the four-year DOD average have a strong statistical relationship with the response.

Next, Figure 25 shows the scatter plots for all pairs of predictor variables displayed in a scatter plot matrix.

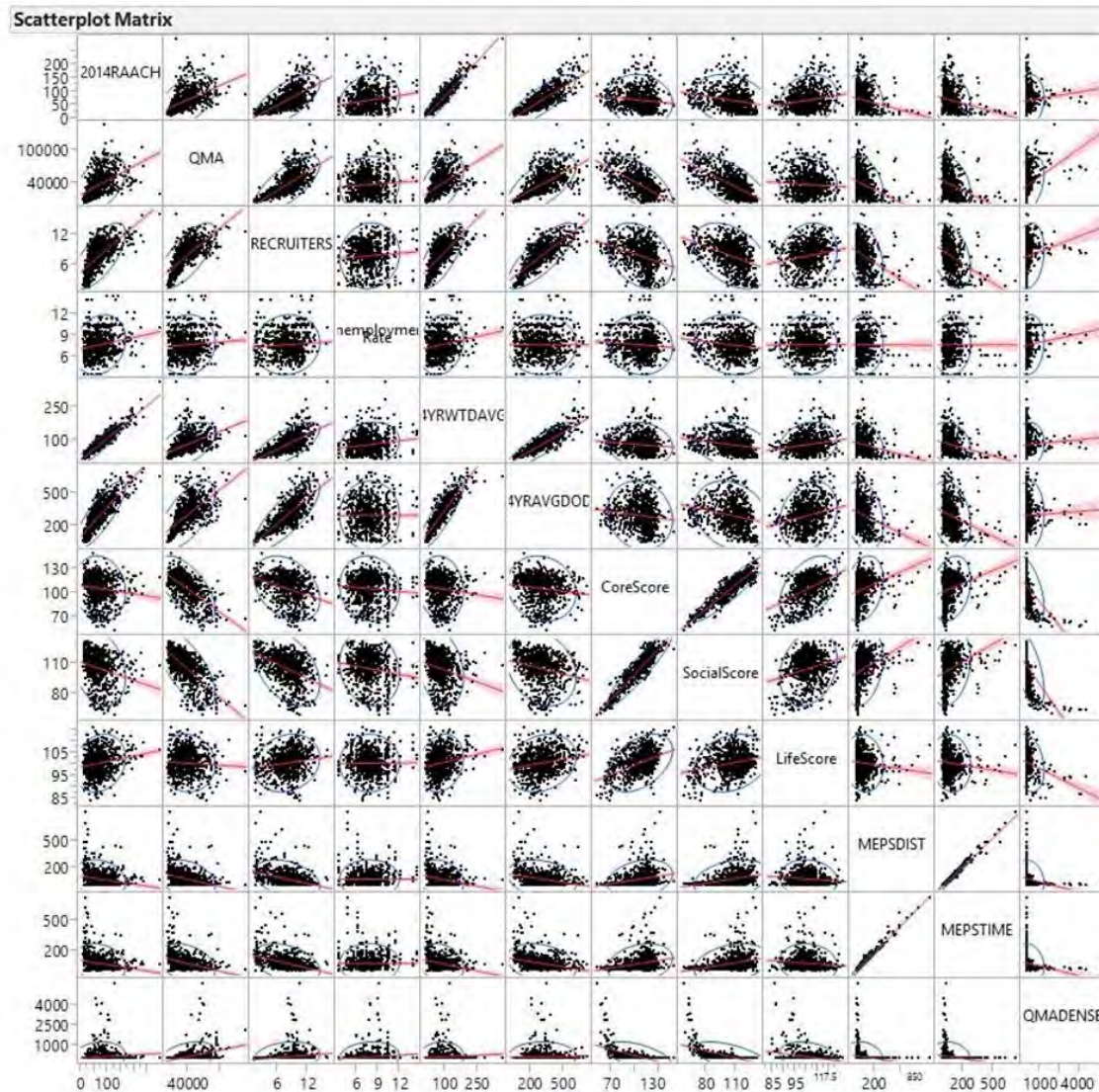


Figure 25. Scatterplot Matrix

The oval shape drawn by the blue line in each scatterplot represents the corresponding bivariate normal density ellipse of the two variables (Klimberg & McCullough, 2013). About 95% of all points would fall inside this ellipse if the two variables are bivariate normally distributed. Additionally, a round ellipse that also does not follow either diagonal signifies that the two variables do not share strong correlation. To aid in identification of correlation, a correlations matrix can be constructed and is shown in Table 8.

Correlations												
	2014RAACH	QMA	RECRUITERS	Unemployment Rate	4YRWTD AVG	4YRAVG DOD	CoreScore	SocialScore	LifeScore	MEPSDIST	MEPSTIME	QMA DENSE
2014RAACH	1.0000	0.4858	0.6751	0.1842	0.9391	0.8310	-0.1236	-0.2353	0.1851	-0.2003	-0.2081	0.0960
QMA	0.4858	1.0000	0.7203	0.0782	0.5073	0.6440	-0.6254	-0.6622	-0.0658	-0.3347	-0.3401	0.3792
RECRUITERS	0.6751	0.7203	1.0000	0.0755	0.7065	0.7987	-0.3359	-0.4192	0.1742	-0.3780	-0.3881	0.1787
Unemployment Rate	0.1842	0.0782	0.0755	1.0000	0.1689	0.0044	-0.0850	-0.1531	0.0095	-0.0027	0.0008	-0.1110
4YRWTD AVG	0.9391	0.5073	0.7065	0.1689	1.0000	0.8747	-0.1159	-0.2205	0.1887	-0.2304	-0.2400	0.0656
4YRAVG DOD	0.8310	0.6440	0.7987	0.0044	0.8747	1.0000	-0.1320	-0.2288	0.2195	-0.2721	-0.2849	0.0354
CoreScore	-0.1236	-0.6254	-0.3359	-0.0850	-0.1159	-0.1320	1.0000	0.9312	0.4475	0.2677	0.2598	-0.5681
SocialScore	-0.2353	-0.6622	-0.4192	-0.1531	-0.2205	-0.2288	0.9312	1.0000	0.2465	0.3114	0.3115	-0.5896
LifeScore	0.1851	-0.0658	0.1742	0.0095	0.1887	0.2195	0.4475	0.2465	1.0000	-0.1041	-0.1141	-0.2621
MEPSDIST	-0.2003	-0.3347	-0.3780	-0.0027	-0.2304	-0.2721	0.2677	0.3114	-0.1041	1.0000	0.9913	-0.2016
MEPSTIME	-0.2081	-0.3401	-0.3881	0.0008	-0.2400	-0.2849	0.2598	0.3115	-0.1141	0.9913	1.0000	-0.1892
QMA DENSE	0.0960	0.3792	0.1787	0.1110	0.0656	0.0354	-0.5681	-0.5896	-0.2621	-0.2016	-0.1892	1.0000

Table 8. Correlation Matrix

Clearly there is strong correlation between MEPS distance and MEPS driving time as well as between LIFE score and CORE score, so these factors as pairs should not be included in the modelling efforts.

APPENDIX D. TRANSFORMED PREDICTOR VALUE PLOTS

When we fit a regression using a transformed response, the predicted values share the scale of this transformation. With our first linear regression, we conduct a power transformation of the response based off the Box-Cox plot. With our second linear regression, we conduct a log transformation of the response. In order to avoid bias, we should transform the predictor values back to the same scale as the original response (2014 Achievement). In JMP, we can create a formula for both of these columns of predicted values and transform them back to the original scale. Then, we use these values a fit an x by y plot against 2014 Achievement. Figure 26 shows these plots for the first linear model on the training set and the validation set.

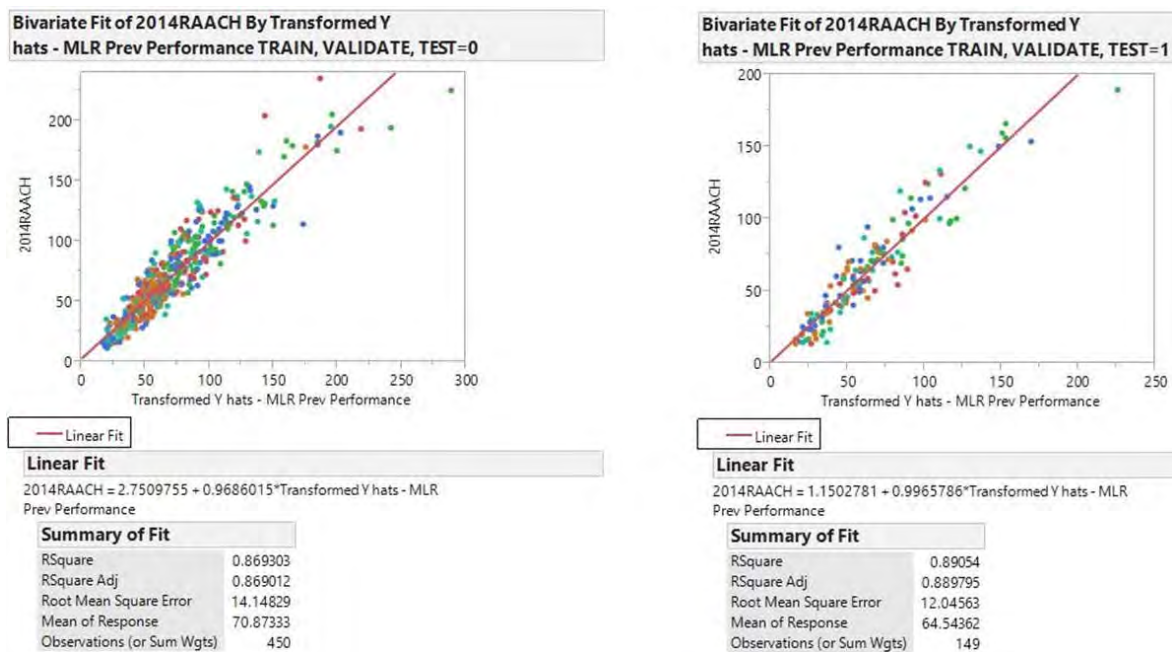


Figure 26. Transformed Predictor Regression Plots—Previous Performance

The resulting R-square value of 0.869 for the training set and 0.89 for the validation set are very close to the original R-square values. The model continues to avoid bias and provide strong predictive power. Next, Figure 27 shows the same x by y

plots for the second linear model, which is the one involving a log transformation of the response.

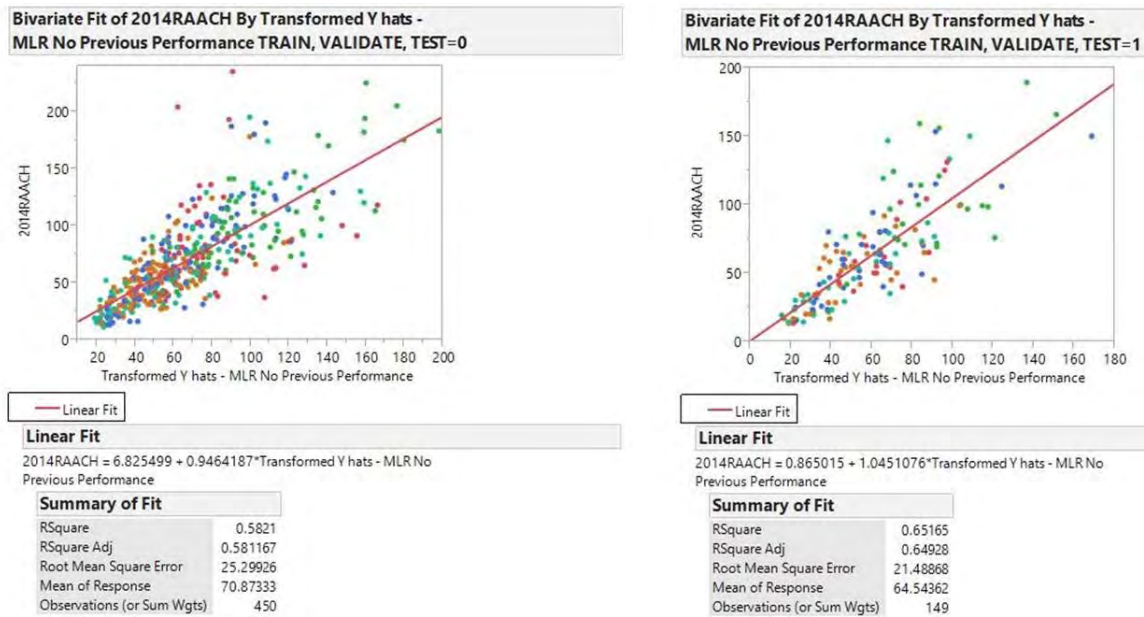


Figure 27. Transformed Predictor Regression Plots—No Previous Performance

The resulting R-squared value on the training set is 0.582, which is lower than the 0.67 received when the predictors are not transformed. However, the R-squared value for the validation set is 0.65, which is very close to the original 0.66 value received when the predictors are not transformed.

LIST OF REFERENCES

- Baird, J. (2014, December 17). CORE Segments. Fort Knox, KY: USAREC G2.
- Claritas. (2015, March 15). *PRIZM*. Retrieved from [www.claritas.com:
http://www.claritas.com/target-marketing/market-research-services/marketing-data/marketing-segmentation/prizm.jsp](http://www.claritas.com/target-marketing/market-research-services/marketing-data/marketing-segmentation/prizm.jsp)
- Clingan, M. L. (2007). *U.S. Army custom segmentation system*. Fort Knox: Center for Accessions Research.
- Dereu, J., & Robbin, J. (1983). *Application of geodemographics to the Army recruiting problem*. Fort Sheridan: Department of the Army.
- Devin, D. (2014). *SAMA methodology validation*. Fort Knox: USAREC G-2.
- Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning data mining, inference, and prediction*. New York: Springer.
- G-5 Public Affairs, USAREC. (2004). *U.S. Army recruiting command history*. Retrieved from <http://www.usarec.army.mil/hq/apa/download/history.pdf>
- Jackson, S. (2015, March). data_CO_clean. College Station, TX: Texas A&M.
- Klimberg, R., & McCullough, B. (2013). *Fundamentals of predictive analytics with JMP*. Cary, NC: SAS Institute.
- Penney, L., Horgen, K., & Borman, W. (2000). *An annotated bibliography of recruiting*. Tampa: United States Army Research Institute.
- SAS Institute. (2013). *JMP 11 Specialized Models*. Cart: SAS Institute.
- USAREC. (2009). *Recruiting operations*. Fort Knox, KY: Headquarters, United States Army Recruiting Command.
- USAREC. (2012). *Recruiting center operations*. Fort Knox: United States Army Recruiting Command.
- USAREC. (2013). *USAREC—About us*. Retrieved from USAREC: <http://www.usarec.army.mil/aboutus.html>
- USAREC G2. (2012). *Segmentation analysis and market assessment (SAMA) reports user guide*. Fort Knox: USAREC G-2.

Williams, T. (2014). *Understanding factors influencing navy recruiting production*.
Monterey: Naval Postgraduate School.

Yu-Wei, C. (2015). *Machine learning with R cookbook*. Birmingham: Packt Publishing.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
Ft. Belvoir, Virginia
2. Dudley Knox Library
Naval Postgraduate School
Monterey, California